Confirmation and Evidence

Inaugural-Dissertation

zur Erlangung der Doktorwürde

der

Philosophischen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität

zu Bonn

vorgelegt von

Jan Sprenger

aus

Köln

Bonn 2008

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn.

Zusammensetzung der Prüfungskommission:

- Prof. Dr. Christoph Horn, Institut für Philosophie
(Vorsitzender)

- Prof. Dr. Andreas Bartels, Institut für Philosophie
(Betreuer und Gutachter)

- Prof. Dr. Rainer Stuhlmann-Laeisz, Institut für Philosophie
(Gutachter)

- apl. Prof. Dr. Hans-Joachim Pieper, Institut für Philosophie
(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: Bonn, 3. Juli 2008.

Erklärung über verwendete Hilfsmittel

Diese Arbeit wurde mit Hilfe des Satzprogramms TEX und der Makrosprache LATEX erstellt. Die verwendete Literatur ist in der Bibliographie aufgelistet. Abbildungen 6.1, 7.1 und 7.2 wurden mit Hilfe des Programms MATLAB erstellt. Bei den anderen Abbildungen wurde der Bildnachweis angegeben.

Anderweitige Hilfen (z.B. Anregungen, die sich Gesprächen mit anderen Forschern verdanken) sind im Text als solche gekennzeichnet.

Danksagung

Mit der Veröffentlichung dieses Buches, meiner Dissertationsschrift, geht für mich nicht nur ein Ausbildungs-, sondern auch ein Lebensabschnitt zu Ende. Drei Jahre lang, von Sommer 2005 bis Sommer 2008, habe ich Literatur gewälzt, ausgiebig diskutiert, Ansatzpunkte zu finden versucht, erste eigene Ideen entwickelt, diese mühevoll ausgearbeitet, zwischendurch am ganzen Projekt gezweifelt, Reisen zu Konferenzen im Ausland unternommen, manches Eigene vorgetragen, vieles Andere angehört, Anregungen daraus gezogen, und vor allem sehr viele spannende Menschen kennengelernt. Aber ich habe in diesen drei Jahren auch viel Musik gehört, Freunde getroffen, gegessen, getrunken, gefeiert und mich manchmal auch entspannt. Diese Ablenkung erwies sich letztlich als ebenso unverzichtbar wie die rein fachlichen Fortschritte. Es erscheint mir somit unmöglich, all jene aufzuzählen, die in der einen oder anderen Weise zum Gelingen des Projekts beigetragen haben. Dennoch möchte ich diejenigen beim Namen nennen, die meine Arbeit in herausragender Weise begleitet haben. Da wäre an erster Stelle mein Betreuer Andreas Bartels, der für mich ein echter Glücksfall war und dem ich für die stets wohlwollende und freundliche Förderung, die geschickte Themenwahl, die Freiheit bei der Bearbeitung, die hilfreichen Ratschläge und die andauernde Unterstützung bei allen Projekten, seien es Stipendienanträge oder Auslandsaufenthalte, zu größtem Dank verpflichtet bin. An unserem Lehrstuhl besaß Jacob Rosenthal stets Zeit und ein offenes Ohr, um Entwürfe gründlich zu lesen, mit mir meine Thesen durchzugehen und sich über (nicht nur) philosophische Themen auszutauschen. Dem Centre for Philosophy of Natural and Social Sciences (CPNSS) an der London School of Economics und dem Tilburg Center for Logic and Philosophy of Science (TiLPS) an der Tilburg University verdanke ich Einladungen zu dreimonatigen, äußerst fruchtbaren Forschungsaufenthalten. Dort traf ich auch auf Stephan Hartmann, der mir sowohl auf fachlicher als auf persönlicher Ebene eine unvergleichliche Förderung zukommen ließ. Das Bundesministerium für Bildung und Forschung bescherte mir durch ein zweieinhalbjähriges Stipendium finanzielle Unabhängigkeit und Forschungsfreiheit; das Cusanuswerk hat dieses Stipendium nicht nur vermittelt, sondern es durch ideelle Förderung auch beträchtlich erweitert. Schließlich danke ich meiner Mutter, die häufig genug zugleich Privatsekretärin, Finanzministerin, Köchin und Umzugshelferin war, einer weiteren, ungenannt bleibenden Person für die Begleitung auf dem Großteil des Weges, sowie meinen Freunden, Bekannten und Kollegen, dank derer die letzten drei Jahre für mich nicht nur in fachlicher Hinsicht erinnerungswürdig bleiben werden.

Zusammenfassung

Die rechtfertigende Kraft sinnlicher Erfahrung und ihr Einfluss auf die Stärke unserer Überzeugungen gehören zu den ältesten Fragen der Philosophie im Allgemeinen und der Erkenntnistheorie im Besonderen. Letztere befasst sich mit der Frage, was menschliches Wissen ausmacht, wie es gebildet wird und wie es mit sinnlicher Erfahrung zusammenhängt. Eine Frage, die häufig in extremer Weise beantwortet wurde: Für Platon konnten sinnliche Erfahrungen lediglich Meinungen (*doxa*) stützen wohingegen wahre Erkenntnis (*epistêmê*) durch die Erinnerung an zeitlose Ideen, die uns angeboren sind, zustande käme. Platons Schüler Aristoteles rehabilitierte hingegen die Erfahrung als eine Basis für allgemeine Behauptungen, z.B. im Rahmen induktiver Schlüsse, die vom Speziellen zum Allgemeinen aufsteigen. Selbst René Descartes, der am Beginn der Neuzeit so sehr die prinzipielle Fehlbarkeit sinnlicher Erfahrung betonte, räumte am Ende der *Meditationes de Prima Philosophia* ein, dass sinnliche Erfahrung viel häufiger zu wahren als zu falschen Überzeugungen führe. Empiristen wie David Hume leugneten sogar, dass es Vorstellungen im menschlichen Geiste gebe, die unabhängig von jeglicher sinnlicher Erfahrung seien. In der Vielfalt der Positionen, die zum Zusammenhang von Erkenntnis und Erfahrung in der Geschichte der Philosophie artikuliert wurden und weitgehende Folgerungen aus der Antwort ableiteten, zeigt sich sowohl der kontroverse Charakter dieser Frage als auch ihre zentrale Funktion in der Erkenntnistheorie. Wie also verhält sich Erfahrung zu Überzeugungen und wie können letztere durch Erfahrung gerechtfertigt werden?

Die Wissenschaftstheorie stellt diese Frage in einer speziellen Hinsicht – nämlich im Hinblick auf das Verhältnis wissenschaftlicher Hypothesen und Theorien zu Daten, die in wissenschaftlichen Experimenten gewonnen werden. In der modernen Wissenschaft nimmt das rigorose Überprüfen und Testen wissenschaftlicher Hypothesen unter Laborbedingungen einen breiten Raum ein, so dass sich aus wissenschaftstheoretischer Perspektive die Frage

stellt, nach welchen Prinzipien Experimente und Beobachtung zum Verwerfen bestimmter Hypothesen führen, während andere Hypothesen durch sie gestützt werden. Diese Frage stellt auch den Kern meines Buches dar. Sie kann allerdings in zweifacher Hinsicht gestellt werden: als Frage nach einem Grund für die rechtfertigende Kraft der Beobachtung überhaupt und als Frage nach einer formalen Theorie der Bestätigung wissenschaftlicher Hypothesen. Die erste Frage wirft das Problem der Begründung induktiver Schlüsse auf: Was gibt uns das Recht, vergangene Erfahrung in die Zukunft zu projizieren? Ich argumentiere jedoch, dass die zweite Frage – die nach den Prinzipien wissenschaftlicher Bestätigung – sich auch sinnvoll stellen und diskutieren lässt, wenn das Induktionsproblem ausgeklammert wird. Dieses Buch befasst somit mit der Explikation der Beziehung zwischen Beobachtungsbelegen und Hypothesen und versucht, die Struktur eines gültigen induktiven Arguments zu erfassen, wobei besonderes Gewicht auf induktives Schließen in der Statistik gelegt wird. Die Antworten sind nicht nur für Bestätigungstheoretiker und Statistiker relevant, sondern wirken sich auch auf mehrere Fragen in der Wissenschaftstheorie aus. So haben zum Beispiel die logischen Empiristen Bestätigbarkeit einer Aussage als Kriterium empirischer Signifikanz vorgeschlagen. Zudem behaupten Argumente für den wissenschaftlichen Realismus häufig, dass manche wissenschaftliche Theorien über einen weiten Zeitraum hinweg hohe Bestätigung genießen und damit die Existenz der von ihr postulierten Größen nahelegen würden. Ebenso ließe sich fragen, welche Rolle der Bestätigungsgrad bei der Dynamik wissenschaftlicher spielt. Eine formale Theorie der Bestätigung stellt somit die Werkzeuge bereit, die es ermöglichen, diese Fragen im Detail zu untersuchen. Davon abgesehen spielen die Begriffe Bestätigung und Evidenz auch außerhalb der Wissenschaftstheorie eine zentrale Rolle, jedoch erfolgt ihre Verwendung häufig in informeller Weise, worunter die Präzision der vorgebrachten Argumente leidet. Ein Beispiel dafür bildet die Diskussion von 'evidence' in der zeitgenössischen Erkenntnistheorie. Ebenso taucht der Evidenzbegriff in den Teilen der Philosophie der Sozialwissenschaften, welche sich mit evidenzbasierten politischen Entscheidungen befassen, an zentraler Stelle auf. All diese Debatten sind jedoch zum gegenwärtigen Zeitpunkt hochgradig informell und ich denke, dass sie stark von den Einsichten profitieren könnten, die die Bestätigungstheorie über den Evidenzbegriff gewinnt.

Die Struktur des Buches ist wie folgt: Ich beginne mit einer Abgrenzung von Induktions- und Bestätigungsproblem (Kapitel eins) und diskutiere im

Anschluss mehrere qualitative Theorien der Bestätigung (Kapitel zwei und drei). Kapitel vier führt die subjektive Interpretation von Wahrscheinlichkeiten als Theorie rationaler Glaubensgrade ein, während Kapitel fünf dies auf die Quantifizierung des Stützungsgrades einer Hypothese anwendet. Der Rest des Buches ist dem induktiven Schließen in der Statistik gewidmet: Kapitel sechs stellt die verschiedenen statistischen Schulen einander gegenüber, während Kapitel sieben Vor- und Nachteile des Evidenzbegriffs in diesen Schulen diskutiert. Kapitel acht fasst schließlich die Ergebnisse zusammen und skizziert einige offene Fragen, für die die gewonnenen Erkenntnisse relevant sind.

Zunächst also der Zusammenhang von Induktion und Bestätigung. David Hume (1777) hat bekanntlich den klassischen Versuch, das Induktionsprinzip durch Verweis auf den vergangenen Erfolg induktiven Schließens zu begründen, widerlegt. Ein solches Argument trägt nämlich selbst induktiven Charakter, so dass die Rechtfertigung zirkulär wird. Eine mögliche Antwort besteht im Bruch mit den Prinzipien des Inferentialismus: Solange die Induktion faktisch einen reliablen Prozess zur Bildung wahrer Meinungen darstellt, brauchen wir sie nicht begründen, um in ihrer Anwendung gerechtfertigt zu sein. Nelson Goodman (1983) weist jedoch darauf hin, dass die entscheidende Frage nicht darin besteht, *ob* Induktion zuverlässig ist, sondern *welche Art* von Induktion zuverlässig ist und zu mehr korrekten als falschen Vorhersagen führt. Das klassische Beispiel hierzu stellt das 'grot'-Problem ('grue' problem) dar – 'grot' bezeichnet ein Prädikat, das auf bereits untersuchte Objekte zutrifft genau dann, wenn sie grün sind, und auf alle anderen Objekte genau dann, wenn sie rot sind. Nach den Prinzipien der Induktion wird die Hypothese 'alle Smaragde sind grün' nun genauso durch die Beobachtung grüner Smaragde in der Vergangenheit gestützt wie die Hypothese 'alle Smaragde sind grot'. Beide Prädikate setzen in der Vergangenheit liegende Beobachtungen in die Zukunft fort, jedoch gelangen sie zu miteinander inkonsistenten Ergebnissen. Nur der induktive Schluss, dass alle Smaragde grün sind, scheint in erkenntnistheoretischer Hinsicht gültig zu sein. Dies wirft ein skeptisches Licht auf Versuche, Prinzipien induktiven Schließens im Rahmen einer formalen Theorie der Bestätigung zu erfassen. In der Tat scheinen wir Zusatzannahmen zu benötigen, z.B. dass manche Prädikate induktiv fortsetzbar sind ('grün'), andere jedoch nicht ('grot'). Diese Annahmen eröffnen den Zugang zum eigentlichen Projekt der Bestätigungstheorie, nämlich die Prinzipien induktiven Schließens aus der Praxis zu extrahieren und die Praxis wiederum durch solche Prinzipien zu korrigieren.

Viele wissenschaftliche Hypothesen sind in den formalen Rahmen einer Logik erster Stufe eingebettet, zum Beispiel die meisten Theorien der klassischen Physik. Andere sind hingegen statistischer Natur, so zum Beispiel die Mendelschen Vererbungsgesetze. Eine Theorie der Bestätigung sollte sich gleichermaßen mit probabilistischen wie mit nicht-probabilistischen Fällen befassen. Hierfür haben sich zwei verschiedene Traditionen entwickelt: die qualitative und die quantitative Tradition, die zumeist in bayesianischer Weise expliziert wird. Ich beginne mit den qualitativen Bestätigungstheorien.

Qualitative Theorien stehen unter einem gewissen Rechtfertigungsdruck, da häufig argumentiert wird, der auf Glaubensgraden aufbauende Bayesianismus sei eine umfassende, allgemeine Theorie induktiven Schließens, die qualitative Ansätze (im Rahmen einer Prädikatenlogik erster Stufe) überflüssig mache. Dem lässt sich jedoch entgegenhalten, dass solche probabilistischen Bestätigungsargumente auf eine große Menge von Fällen in der Wissenschaftsgeschichte nicht anwendbar sind, weil die *Struktur* der induktiven Argumente nicht korrekt erfasst wird und die Argumente häufig keine Glaubensgrade verwendeten (vgl. Glymour 1980a). Zum Beispiel finden sich Glaubensgrade weder in Eddingtons Bestätigung der Allgemeinen Relativitätstheorie noch in der Bestätigung der Newtonschen Theorie durch die erfolgreiche Vorhersage der Wiederkehr des Halleyschen Kometen. Dies rechtfertigt die Untersuchung qualitativer Ansätze. An einer Reihe von Beispielen (unter ihnen das berühmte Rabenparadoxon) lässt sich ferner deutlich machen, dass die Bestätigungsrelation als dreistellige Relation aufgefasst werden sollte, die neben Hypothese und Beobachtungsbelegen überdies noch eine Menge an Hintergrundannahmen enthält.

Zwei grundverschiedene Ansätze prägen die qualitative Bestätigungstheorie: induktivistische Ansätze, zu denen Hempels Erfüllungskriterium ('satisfaction criterion') gehört und die hypothetisch-deduktiven Ansätze, die sich auf Poppers (1963) Modell von Vorhersage, Test und empirischer Bewährung berufen. Zwei hauptsächliche Kritikpunkte lassen sich gegen den Hempelschen Ansatz einwenden: Zum einen leidet er unter einer beträchtlichen Zahl technischer Schwierigkeiten. Zum anderen diagnostiziert Hempel [1945] (1965) zwar korrekt, dass das Rabenparadoxon lediglich paradox *erscheint* und dass stillschweigendes Hinzufügen zusätzlichen Hintergrundwissens für diese paradoxe Erscheinung verantwortlich ist. Insbesondere sind induktive Schlüsse nicht *monoton* – wenn weiteres Hintergrundwissen zu den Prämissen hinzugefügt wird, kann die Gültigkeit des Schlusses verloren gehen. Dies unterscheidet induktive Schlüsse wesentlich von deduktiven Schlüssen. Jedoch gelingt es Hempel nicht, diese Erkenntnis in sein eigenes Bestätigungskriterium zu integrieren, welches eine monotone Theorie induktiven Schließens darstellt. Somit schlägt Hempels Kriterium nicht nur im Falle des Rabenparadoxons fehl, sondern immer dann, wenn das Hintergrundwissen substantiell erweitert wird.

Das Scheitern des Hempelschen Kriteriums motiviert eine Untersuchung der zweiten, hypothetisch-deduktiven Tradition. Hiernach machen Hypothesen Voraussagen mit der Hilfe von Hintergrundannahmen, und Bestätigung besteht darin, dass diese Vorhersagen tatsächlich eintreten. Genauer formuliert, folgt die Evidenz oder der Beobachtungsbeleg logisch aus der Hypothese und den Hintergrundannahmen. Dieser Ansatz erfasst ein Grundmuster experimenteller Praxis in der Wissenschaft, jedoch gelingt es ihm zumindest in seiner elementaren Form nicht, die *Relevanz* eines Beobachtungsbelegs für eine Hypothese einzufangen: Wenn ein Beleg E eine Hypothese H im hypothetisch-deduktiven Sinn bestätigt (relativ zu Hintergrundannahmen K), so bestätigt E auch die Konjunktion von H mit einer nahezu beliebigen Hypothese X. Dies ist klarerweise unerwünscht, da E in aller Regel nicht relevant für X ist. Mehrere Versuche, dieses Problem zu lösen, scheiterten, und erst in den neunziger Jahren des letzten Jahrhunderts konnten zufriedenstellende Ergebnisse (Schurz 1991, Gemes 1993) entwickelt werden. Allerdings gibt es auch hier einen Aspekt, der nicht gelöst werden konnte, nämlich die Bestätigung von Hypothesen, die aus mehreren Einzelhypothesen bestehen, welche konjunktiv miteinander verknüpft sind. Die gängigen Modelle hypothetisch-deduktiver Bestätigung implizieren häufig, dass, wenn E_1 H_1 bestätigt und E_2 H_2 bestätigt, auch $H_1.H_2$ durch $E_1.E_2$ bestätigt wird (alles relativ zu K). Ich zeige anhand eines der Wissenschaft entnommenen Beispiels, dass solche induktiven Schlüsse nicht im Allgemeinen gültig sein können. Insbesondere vernachlässigt dieses Schema, dass in vielen Fällen *Instanzen* einer Hypothese beobachtet werden müssen, um diese zu bestätigen. In meinem eigenen Lösungsvorschlag verbinde ich daraufhin die bestätigende Kraft von Instanzen einer Hypothese mit einem falsifikationistischen Kriterium in der Popperschen Tradition von Mutmaßungen und Widerlegungen. Das neue, falsifikationistische Kriterium löst nicht nur die Probleme evidentieller Relevanz ebenso wie die konjunktiv zusammengesetzter Hypothesen, es ist überdies auch noch einfacher formulierbar als seine Konkurrenten. Mithin stellt es das meiner Meinung nach am weitesten fortgeschrittene und präziseste Kriterium für qualitative Bestätigung dar.

Häufig geht es in der Wissenschaft nicht nur darum, einfache Hypothesen, sondern Hypothesenkomplexe oder ganze Theorien durch eine Datenmenge zu bestätigen. Darüber hinaus fehlen häufig externe, theorieunabhängige Hintergrundannahmen, gegen die sich eine Theorie testen ließe – dies lässt sich anhand Kuhns (1962) Bemerkung zur Theorieabhängigkeit von Beobachtungen illustrieren. Ein Modell, welches die Bestätigung einer ganzen Theorie auf Bestätigungsrelationen innerhalb dieser Theorie zurückführt, wird durch Clark Glymours (1980a) Bootstrap-Bestätigung gegeben: Theorien werden durch deduktive Schlüsse von Beobachtungsbelegen und Teilen der Theorie auf andere Teile der Theorie bestätigt. Jedoch lassen sich dagegen mehrere technische Einwände aufstellen (Christensen 1983, 1990), welche Glymours ursprüngliches Bestätigungsmodell widerlegen. Ich argumentiere, dass zwei prinzipielle Antworten möglich sind: entweder fasst man Bootstrap-Bestätigung als Modell der *Kohärenz* zwischen Theorie und Belegen (anstelle von Bestätigung) auf, oder das Hempelsche Erfüllungskriterium, welches der Bootstrap-Bestätigung zugrunde liegt, wird durch ein anderes Kriterium ersetzt. Hier zeige ich, dass das falsifikationistische Kriterium in der Lage ist, die Bootstrap-Bestätigung zu retten und – für eine derart modifizierte Form der Bootstrap-Bestätigung – Christensens Einwände zurückzuweisen.

Wie alle Kriterien qualitativer Bestätigung hat auch das falsifikationistische Kriterium und die falsifikationistische Version der Bootstrap-Bestätigung mit dem Duhem-Quine-Problem zu kämpfen. Duhem (1914) argumentierte, dass jeder Test einer empirischen Hypothese ein Konglomerat an Hilfshypothesen benötige und ein negatives Ergebnis nur so sehr gegen die getestete Hypothese spreche wie man Vertrauen in die verwendeten Hilfshypothesen habe. Daraus folgerte Quine (1961), dass es keine Bestätigung einzelner Hypothesen durch Beobachtungen geben könne – solche Beobachtungen würden zwar unser Theoriennetzwerk als Ganzes beeinflussen, jedoch niemals einzelne Hypothesen. Nichtsdestoweniger glaube ich, dass man Duhems These zustimmen kann, ohne Quines Folgerung zu teilen. Zwar wirkt sich ein Beobachtungsbeleg nicht nur auf eine einzelne Hypothese, sondern auch auf die benutzten Hilfshypothesen aus, jedoch zumeistin unterschiedlichem Maße. Wenn also eeine Hypothese in einem Test scheitert, werden alle beteiligten Hilfshypothesen unterminiert, aber der Grad der Unterminierung zeigt an, welche Hypothesen davon stärker betroffen sind als andere.

Um diese unterschiedlichen Grade der Bestätigung oder Unterminierung einzelner Hypothesen einzufangen, wird eine quantitative Bestätigungstheorie benötigt. Diese soll im Folgenden skizziert werden. Die grundlegende Idee bildet dabei, Bestätigung als Anstieg des rationalen Glaubensgrades in eine Hypothese zu modellieren. Dafür brauchen wir ein formales Modell, das die Veränderung von Glaubensgraden abbildet und Rationalitätskriterien für Systeme von Glaubensgrade formuliert. Ein solches Modell wird durch die Wahrscheinlichkeitstheorie gegeben. Wenn die Stärke eines Glaubensgrades als Urteil über die Fairness hypothetischer Wetten verstanden wird, dann sind genau jene Systeme von Glaubensgraden rational, in denen sich keine Systeme von Wetten konstruieren lassen, die einer Seite einen sicheren Vorteil versprechen. Die Dutch-Book-Argumente zeigen, dass dies genau jene Systeme von Glaubensgraden sind, die den Axiomen der Wahrscheinlichkeitstheorie genügen. Aufbauend auf diesem subjektiven, bayesianischen Verständnis von Wahrscheinlichkeiten als Stärke eines Glaubensgrades lässt sich dann der Bayesianismus als quantitative Bestätigungstheorie formulieren: Bestätigung besteht im Anstieg des rationalen Glaubensgrades. Die entscheidende Frage ist jedoch, wie die Stärke der Bestätigung expliziert wird, und es stellt sich heraus, dass mehrere Probleme in der bayesianischen Bestätigungstheorie davon abhängen, welches Bestätigungsmaß gewählt wird. Mit Hilfe plausibler Adäquatheitskriterien lässt sich eine Reihe von Maßen ausschließen. Die verbleibenden Maße explizieren dann zwei verschiedene Begriffe von Bestätigung und induktiver Stützung: zum einen Bestätigung als Verallgemeinerung logischer Folgerung (und Bestätigungsgrade als Stärke eines induktiven Arguments), zum anderen Bestätigung als Einfluss eines Beobachtungsbelegs auf den epistemischen Status einer Hypothese. Meines Erachtens hängt es stark vom jeweiligen Anwendungskontext ab, welches Maß bevorzugt werden sollte. Diese Kontextabhängigkeit ist in der bestehenden Literatur vernachlässigt worden und überträgt sich auch auf Versuche, das Problem alter Evidenzen ('problem of old evidence') zu lösen.

Die bayesianische, wahrscheinlichkeitstheoretische Auffassung von Bestätigung ermöglicht es in natürlicher Weise, quantitative Bestätigungstheorie auf statistische Regelmäßigkeiten anzuwenden. Dies ist in der Tat die interessanteste Anwendung der quantitativen Bestätigungstheorie, da statistische Methoden mehr und mehr Platz in den empirischen Wissenschaften einnehmen. Zum Beispiel hat gerade in den angewandten Wissenschaften die Anzahl der verfügbaren Daten in den letzten Jahrzehnten so stark zugenommen, dass statistische Analysen absolut unverzichtbar geworden sind. Ein Problem der bayesianischen Statistik stellt dabei dar, dass sie keine komplett objektive Schule des induktiven Schließens liefern kann, da sie auf subjektiven Glaubensgraden und Urteilen aufbaut. Wissenschaftler möchten jedoch häufig die Ergebnisse ihrer Forschungen in objektiver Weise zusammenfassen und keinen Spielraum für subjektiven Dissens lassen. Zwar gibt es Abarten des Bayesianismus (wie das Prinzip der maximalen Entropie), welche einen solchen Dissens ausschließen, aber dagegen lassen sich eigene Einwände machen – zum Beispiel, dass subjektive Expertise vernachlässigt wird und dass Mangel an Information mit Hilfe bestimmter Wahrscheinlichkeitsverteilungen dargestellt wird.

Dass der Bayesianismus eine im Ganzen subjektive Theorie induktiven Schließens darstellt, impliziert nicht, dass er nicht in der Lage wäre, eine weitgehend objektive Theorie statistischer Evidenz zu liefern. Bayesianische Evidenzmaße wie Likelihood-Quotienten genügen Birnbaums (1962) Likelihood-Prinzip, dass alle statistische Evidenz als Funktion der Wahrscheinlichkeiten des beobachteten Ereignis unter den verschiedenen Hypothesen darzustellen ist. Dies schließt subjektive Faktoren in der Regel aus. Obwohl das Likelihood-Prinzip sich auf zwei plausible elementare Prinzipien zurückführen lässt, wird es in der statistischen Praxis oft bewusst verletzt. Viele Statistiker glauben nämlich, dass das Charakteristikum eines gültigen induktiven Schlusses in der Reliabilität der verwendeten Methode bestehe (Mayo und Spanos 2006). Darauf basiert die Test- und *Fehlerstatistik*: je geringer die Wahrscheinlichkeit, dass sich eine Entscheidungsprozedur irrt, desto eher sollten wir das Ergebnis akzeptieren. Der Hauptvorteil dieser Methode liegt darin, dass der Subjektivitätsvorwurf, der gegen den Bayesianismus erhoben wird, hier nicht greift.

Obwohl die Fehlerstatistik auf anscheinend plausiblen Prinzipien beruht, lässt sich zeigen, dass fehlerstatistische Evidenzmaße wie Fehlerwahrscheinlichkeiten, Signifikanzniveaus und p-Werte gravierende Probleme aufweisen. Von eher technischen Einwänden abgesehen reagieren sie nämlich auf einen Wechsel des experimentellen Designs, d.h. die Stärke der Evidenz wird dadurch beeinflusst, ob der Wissenschaftler den Plan zur Durchführung des Experiments korrekt wiedergegeben hat. Solche Faktoren lassen sich jedoch nicht durch Replikation eines Experiments überprüfen, so dass sie nicht in ein Evidenzmaß eingehen sollten, das quantifiziert, wie sehr eine Hypothese durch Daten gestützt wird. Das bedeutet natürlich nicht, dass experimentelles Design komplett irrelevant wäre – es spielt lediglich keine Rolle in der post-experimentellen Beurteilung der Stärke der beobachteten Evidenz. Ich argumentiere ferner, dass der Evidenzbegriff *komparativ* aufgefasst werden sollte, d.h. nicht als Evidenz für eine einzelne Hypothese oder im Vergleich zu einer unspezifischen Alternative. Statt dessen sollte ein fruchtbarer Evidenzbegriff immer zwei bestimmte Hypothesen miteinander vergleichen. All dies impliziert, dass die bayesianische Explikation statistischer Evidenz als Grundlage faktischer Entscheidungen der fehlerstatistischen Explikation vorzuziehen ist.

Abschließend möchte ich einige allgemeine Schlussfolgerungen ziehen und offene Fragen skizzieren. Erstens lassen die grundlegenden Vorteile des bayesianischen Ansatzes gegenüber seinem fehlerstatistischen Konkurrenten darauf schließen, dass keine Theorie induktiven Schließens zugleich vollständig objektiv und universell anwendbar ist und dabei den Evidenzbegriff in vernünftiger Weise expliziert. Qualitative Kriterien wie das falsifikationistische Kriterium scheiden aus, weil ihnen die quantitative Dimension fehlt, die für einen Großteil moderner statistischer Anwendungen unverzichtbar ist. Jedoch stellen sie wichtige Hilfsmittel dar, um Fälle von Bestätigung in der Wissenschaftsgeschichte zu modellieren. Der den partiell subjektiven Charakter induktiver Schlüsse anerkennende Bayesianismus scheint im Großen und Ganzen unser bester Kandidat zu sein. Das impliziert jedoch auch, dass sich manche Fragen induktiven Schließens – insbesondere dann, wenn die Datenlage dünn und die Evidenz nicht stark ist – nicht eindeutig klären lassen und auf subjektive Expertise angewiesen bleiben.

Zweitens bietet es sich an, die Debatte zwischen Bayesianern und Fehlerstatistikern auf allgemeine erkenntnistheoretische Fragen auszuweiten. Der erkenntnistheoretische Reliabilismus behauptet, dass Subjekte gerechtfertigte Meinungen unterhalten, insofern diese von einem zuverlässigen kognitiven Kausalprozess erzeugt werden. In derselben Weise rechtfertigen die Methoden der Fehlerstatistik einen induktiven Schluss dadurch, dass die ihn erzeugende Methode eine hohe Reliabilität besitzt. Die Reliabilität einer verwendeten Methode genügt jedoch nicht, um eine *Handlung* im Sinne der Entscheidungstheorie zu rechtfertigen, da die anfängliche Wahrscheinlichkeitsverteilung stark variieren kann. Diese kann die Fehlerstatistik aber nicht liefern. Für den erkenntnistheoretischen Reliabilisten stellt sich somit die Aufgabe, den Begriff der reliablen, zuverlässigen Prozedur in nicht-trivialer Weise zu explizieren.

Drittens verdient Statistik als eigenständige wissenschaftliche Disziplin mehr Aufmerksamkeit in der Wissenschaftstheorie als sie bisher erhalten hat. Häufig als rein formale (Hilfs-)Wissenschaft aufgefasst, ist sie in Wirklichkeit äußerst eng mit den empirischen Wissenschaften verbunden, aus denen sie auch stammt und wo eine hohe Anzahl von Forschern sich der statistischen Analyse verschrieben hat. Sie nimmt eine faszinierende Mittelstellung zwischen mathematischer Theorie und empirischer Anwendung ein. Daneben zeigen die originellen Beiträge von Statistikern und empirischen Wissenschaftlern wie Allan Birnbaum, James O. Berger und Richard Royall die Bedeutung der Statistik für die Erkenntnis der Grundlagen induktiven Schließens und somit auch für die Wissenschaftstheorie als Ganzes.

Dieses Buch lässt eine Reihe von interessanten Themen (wie die Rolle der nicht-parametrischen Statistik oder die Analyse statistischer Modelle) aus. Nichtsdestotrotz erhoffe ich mir, dass es die Prinzipien induktiven Schließens in verständlicher und gewinnbringender Weise darstellt und die Leser aus der Auseinandersetzung mit meinen Thesen ihren Nutzen ziehen können.

Contents

Appendices

Introduction

The question how experience acts on our beliefs and how beliefs are changed in the light of experience is one of the oldest and most controversial questions in philosophy in general and epistemology in particular. Hearing an approaching thunderstorm leads to the expectation that it will soon rain. Seeing a friend in a Greek restaurant makes us believe that she likes Greek food. Feeling a permanent scratchiness in the throat triggers the fear to have caught a cold. Epistemology has always been concerned with the question how knowledge is formed and how the impact of experience on the formation of knowledge can be described. For Plato, sensations and experience can only form mere opinions (*doxa*) whereas true knowledge (*epistêmê*) is acquired by perceiving and recognizing the timeless forms and ideas with which we were born and which are buried in our souls (*anamnesis*). Plato's student Aristotle, however, rehabilitated experience as the basis for the assertion of general claims, e.g. in inductive generalizations. Later on, at the beginning of the modern era, René Descartes pointed out the fallibility of sensory experience, but at the very end of his *Meditationes de Prima Philosophia*, he also admitted that sensory experience forms true beliefs much more often than false beliefs. Empiricists as David Hume even went so far to claim that all ideas in the human mind and all justification of human beliefs ultimately go back to sensory impressions. The diversity of positions points to a substantial controversy as well as to the central role of this question in epistemology. How does empirical justification work and how can the relation between experience and justified belief be spelled out?

Philosophy of science has replaced this question by the more specific enquiry how results of experiments act on scientific hypotheses and theories. Since the Renaissance, science has developed a successful method to acquire knowledge about the world by subjecting hypotheses to systematic experimental scrutiny, eventually resulting in rejection, acceptance or modification of a hypothesis. This method of rigorous experimental testing has proved to be incredibly fruitful and successful. Karl Popper (1963) even went so far to claim that subjecting conjectures to empirical test and potential falsification is the hallmark of scientific method as opposed to pseudo-science and dogmatic belief. Still, the old question of empirical justification remains open and triggers a lot of related questions: How do experimental observations act on abstract theories and why do we maintain some theories while discarding others? Why do some scientific models survive and why do others fail to withstand experimental tests? How can abstract theories be connected at all to concrete experiments? All those queries can be subsumed under two general questions: First, what is our reason to accept the justifying power of experience and more specifically, scientific experiments? Second, how can the relationship between theory and evidence be described and under which circumstances is a scientific theory confirmed by a piece of evidence? The book focuses on the second question and maintains that the search for formal criteria for confirmation and disconfirmation is meaningful even if the answer to the first question is left open. We would like to explicate the relationship between theory and evidence and to capture the structure of a valid inductive, ampliative argument. Special attention is paid to statistical applications that are prevalent in modern empirical science.

This project is not only relevant for issues of confirmation and induction, it actually touches several areas in philosophy of science and beyond: In the first half of the twentieth century, the logical positivists proposed confirmability as a criterion for distinguishing empirically significant statements from metaphysical nonsense. Arguments for scientific realism often stress that some theories remain well-confirmed over a large time horizon, thus giving us a reason to believe in their truth and in the existence of the quantities which they posit. The dynamics of scientific theories can be studied as a function of the confirmatory status of the theories. For instance, it might be interesting to ask how much disconfirmation is imposed on a theory by persistent anomalies or whether a series of disconfirming observations is able to trigger a scientific revolution. Formal theories of confirmation open the way to making such arguments explicit. Apart from that, the concepts of confirmation and evidence are present in a wide range of debates outside philosophy of science, but their vague and informal use often blurs the controversy. Contemporary epistemology which has recently developed much interest in the concept of evidence gives a salient example. There are other fields of application, too: In the philosophy of the social sciences, evidence-based decisions and evidence-based policy-making are a major issue. Even in practical philosophy, we might (given some realist intuitions) be interested in evidence for moral claims or in the confirmation of moral hypotheses. Those debates are largely informal and I believe that they could benefit a lot from insights into the nature of scientific evidence.

After an introductory chapter about the link between confirmation and induction, the project starts with discussing qualitative accounts of confirmation in first-order predicate logic (chapter two). Two major approaches, the Hempelian satisfaction criterion and the hypothetico-deductivist tradition, are contrasted to each other. This is subsequently extended to an account of the confirmation of entire theories as opposed to the confirmation of single hypothesis (chapter three). Then quantitative theories of confirmation and the Bayesian account of confirmation (chapter five) are explained and discussed on the basis of a theory of rational degrees of belief (chapter four). After that, I present the various schools of statistical inference and explain the foundations of these competing schemes (chapter six). Chapter seven revolves around the concept of statistical evidence and tries to resolve and to decide the dissent between the various statistical schools. Finally, chapter eight summarizes the results.

Chapter 1

Induction and Confirmation

Science aims at describing, explaining and predicting the real, empirically given world. Physical theories describe the behavior of a harmonic oscillator, biological theories explain photosynthesis, astronomical theories predict solar eclipses. Therefore scientific theories have to stand up against experience which can support, question or refute them. Theories which survive experimental tests and are in agreement with observable phenomena have a better reputation than those which fail to agree with experience. If a particular astronomical theory successfully predicts solar eclipses, we prefer it to a rival theory which does not enjoy that success. The empirical evidence seems to speak in favor of the first and against the second theory. This book tries to characterize how experiments and empirical observation on the one side support or undermine theories and hypotheses on the other side. More precisely, we examine how observations support and undermine scientific theories, how they lead to their endorsement and how they guide us towards their rejection. We would like to discover the mechanics of *confirmation*, *evidence* and *inductive support*, and this book is devoted to studying them.

However, there is a fundamental problem which precedes our investigations: When we prefer a particular theory to another, when we accept one theory but reject another, we express expectations about their future success. But is it all possible to infer from past experience (e.g. observations that we have made) to rational expectations about future success? How do we justify our expectations that ravens which we will observe in the future will be black, like the ravens we encountered in the past? This is the *problem of induction.* It is intimately tied to the confirmation of scientific theories, but, as we will see, questions about induction are not identical to questions about confirmation. The first chapter of this book thus introduces the problem of induction in more detail and illuminates its relationship to the main topic of this book – the confirmation of scientific theories.

1.1 The classical problem

Classical logic is concerned with truth-preserving inferences – inferences that preserve the truth in passing from the antecedens to the consequens, from the premises to the conclusion. This kind of inference, henceforth called *deductive inference*, plays an outstanding role both in scientific and everyday reasoning. Assume that we know that Alice will come to Bob's party if he invites her. Moreover, we know that Bob invites Alice to his party. This entitles us to conclude that Alice will come to Bob's party. This inference was truth-preserving – the truth of the conclusion (Alice will come) was derived from the (alleged) truth of the premises (that Bob invites Alice and that she positively responds to an invitation). Similarly, deductive inferences play an outstanding role in science. Mathematical inferences are truth-preserving – given a set of fundamental axioms (e.g. the Zermelo-Fraenkel-axioms), mathematical theorems are valid formulae. Since mathematical methods are indispensable in empirical science as physics, chemistry or economics, such truth-preserving, deductive inferences are widespread in science. Moreover, our everyday reasoning often relies on deductive inferences, as we have seen in the Alice/Bob-case. However, they do no exhaust the inferences which we make in science. For instance, we might be interested in the color of ravens. We make an observational field study, observing a hundred ravens all of which are black. Now, it seems to be the most natural thing to generalize these observations and to conclude that the next raven which we observe will be black. But this inference, however justified it may be, it certainly not deductively warranted. That we observed one hundred black ravens does not have any logical consequences for the color of any other raven on Earth. Such *inductive* inferences are ampliative – the content of the conclusion goes beyond the content of the premises. By contrast, deductive inferences are non-ampliative – everything that the conclusion might assert is already contained in the premises.

Inductive inferences often occur in science – not only when we examine natural kinds as ravens and observe their color but every time we generalize from experience. Newton's law of gravitation, for instance, makes a claim about the gravitational forces acting between two point masses m_1 and m_2:

$$F = \gamma \frac{m_1 m_2}{r^2} \tag{1.1}$$

where γ is the gravitational constant and r the distance between the point masses. The problem of idealization put aside (proper point masses with zero extension do not exist), Newton's law of gravitation cannot be inferred deductively from single experiments or observations, even not approximately so. Even if we had experimental data where any confounding factors could be ruled out and where our observations were in agreement with Newton's law, we would merely have instances of the gravitational law, but we could not derive it: For Newton's law of gravitation is a universal hypothesis whose domain are all (idealized) point masses in the universe, in the past as well as in the future. Clearly, it is impossible to derive or to prove such a strong claim just out of raw data. It is logically possible that Newton's law holds true of all situations in the past but that it fails to hold of future situations. Actually, this is typical of scientific theorizing – we form strong and general conjectures on the basis of paradigmatic examples and hope that they will hold true although we are not able to demonstrate them in a rigorous way. But in principle, space for doubt will always remain – almost no scientific hypothesis is conclusively established.

So what did justify the NASA astronomers' confidence in Newton's law of gravitation when calculating the path of the Voyager space probes through space? Clearly, the empirical content of Newtonian (or relativistic) mechanics was much stronger than what was warranted by observation. Nevertheless we think that the NASA astronomers acted rationally when they trusted in Newton's gravitation law for calculating the orbit of the space probes and determined a time point where the space probes would take off. Indeed, merely to rely on claims that are proven in a strict sense seems to demand too much from science. It would be too conservative a strategy to tackle real problems. The correct predictions which Newton's law of gravitation made in the past *justify* our expectation that it will correctly predict the behavior of the Voyager space probes. This kind of reasoning exemplifies the *principle of induction*: past successes of a hypothesis justify the expectation that the hypothesis will succeed in the future, too, although they do not logically entail any future success. This principle sounds very plausible and in practice, it is a basic pillar of all human action and theorizing. Nonetheless, in his 'Enquiry Concerning Human Understanding' (1777), David Hume revealed a lacuna in inductive reasoning and put into doubt the justification of the inductive principle itself. When we presume that the past success of scientific hypotheses transfers to their future success, we assume that nature is uniform in time and that the future will resemble the past. Hume points out that this is itself an inductive inference. Usually, it is justified on pragmatic grounds – since the principle of induction was successful in the past, we should apply it in the future, too. For instance, science largely relies on the principle of induction, and since science is arguably a very successful human activity, we are justified to apply the principle of induction in the future, too. But this defense is viciously circular – for justifying the principle of induction we make an inductive inference: the past success of the induction principle is supposed to justify its future application. Hence, we end up in a circle. Thus it is unclear whether experience really tells us what we should believe:

> "To say [an inference from past to future instance] is experimental, is begging the question. For all inferences from experience suppose, as their foundation, that the future will resemble the past, and that similar powers will be conjoined with similar sensible qualities. If there be any suspicion that the course of nature may change, and that the past may be no rule for the future, all experience becomes useless, and can give rise to no inference or conclusion. It is impossible, therefore, that any arguments from experience can prove this resemblance of the past to the future; since all these arguments are founded on the supposition on that resemblance."[1]

Thus, Hume's scepticism towards scientific reasoning is actually very deep and concerns more than the trivial claim that scientific results cannot be known with certainty. We have no logical basis and no convincing reason to place confidence in any scientific prediction based on past and present observations. The space of observations consistent with past observations is endless – when we observe one hundred black ravens, the next 100 ravens could be white, red, or whatsoever. Any defense based on past experience and past success would itself employ the principle of induction and the uniformity of nature in time and thus beg the question. According to Hume, any inductive inference is ultimately a matter of custom and habit – the past

[1] Hume 1777, 37-38.

success of inductive inference creates the habit to trust in inductive inference, but ultimately, we cannot justify why such inferences are valid.

In his answer, Hume gives a psychological explanation why we apply the principle of induction and not an epistemically convincing reason to do so. However, the search for such an ultimate justification might be misguided. From the point of view of an epistemic reliabilist, we remain justified to apply the principle of induction as long as induction is *factually* a reliable method to generate successful predictions.[2] The reliabilist requires no inferential justification of a reliable method – all that is required is that "inductive arguments lead on the whole to true opinions"[3]. Moreover, in 'Fact, Fiction and Forecast', Nelson Goodman (1983) draws an analogy between inductive and deductive inference. We do not judge the soundness of a deductive inference by appeal to a superordinate principle, but by showing that the inference conforms to the *rules* of deductive inference. Therefore, "to justify a deductive conclusion [...] requires no knowledge of the facts it pertains to"[4]. These rules of deductive inference are in turn justified by their conformity to the practice of making deductive inference. This circle is a virtuous one, or so Goodman argues: Rules and inferences are brought in agreement each other by mutual adjustment. Rules are extracted from the practice of making inferences, and inferential practice has to conform to the rules.

This book cannot discuss proposals to *solve* the problem of induction. To recall, we wanted to focus on another question, namely: How does the discipline of confirmation theory stand to the problem of induction? Here I believe that Goodman is right in an important respect: The proper problem of induction does not consist in justifying an inductive inference by appeal to a superordinate principle or meta-principle. Instead, it deals with defining the difference between valid and invalid inductive inference and in finding out which rules capture a sound inductive inference, or so Goodman argues.[5] Confirmation theory is not 'inductivist' in the sense that it makes any assertion about the validity of the induction principle, rather, it sets up formal

[2] Credits to Thomas Grundmann for calling my attention to this possibility.

[3] Ramsey [1926] 1978, 99.

[4] Goodman 1983, 63.

[5] One might add that ultimately, it will also be necessary to prove soundness and completeness theorems for a logic of inductive inference, as it has already been done for deductive logic, in order to show that valid inferences are precisely those inferences that conform to the rules of induction.

models in which our inductive practice can be formalized and analyzed.[6] There is a division of labor between answering Hume's challenge and structuring inductive practice, and we focus on the second task. However, as we will soon see, this requires a specification of the conditions under which confirmation theory is really independent of the induction problem as such.

For instance, the observation of one hundred black ravens seems to better support the hypothesis that all future ravens are black than the hypothesis that all future ravens are white. We notice that this assertion employs a nontrivial *inductive principle*, namely the assumption that nature is uniform in time. For this very reason, we expect the next observed raven to be black, too. Our task is now *explicative*: A vague and imprecise concept – 'confirmation' or 'valid inductive inference' – is to be replaced by a precise, tractable, fruitful and simple concept that is as similar as possible to the old, imprecise predecessor.[7] Our rules wound then distinguish between valid inductive inferences (as the step from the observation of many black ravens to 'all ravens are black') and invalid inductive inferences (as the step from the same observation to 'all ravens are white').[8]

The discipline of extracting rules from practice, putting them into a consistent calculus and proving theorems for that calculus is often called *inductive logic* (Fitelson 2005, Hawthorne 2008). Like deductive logic, it is neutral with respect to potential applications in the real world. While deductive logic develops a formalism that discerns *truth-preserving* inferences, inductive logic studies inferences where the truth of the premises *indicates* the truth of the conclusions without guaranteeing it. Since the main application of induction is the confirmation of scientific hypotheses, inductive logic often figures as *confirmation theory*, too. In practice, the distinction between inductive logic and confirmation theory concerns less the subject of inquiry than the various research traditions. Confirmation theorists are interested in applications to real science whereas inductive logicians investigate the relationship between the validity of an inductive inference and inductive rules and try to prove soundness and completeness theorems. But finally, they would like to answer the same question: how do we explicate valid inductive inferences? Answering that question involves a tradeoff between normative and descriptive adequacy: On the one hand, if inductive practice in science

[6] This point was brought to my attention by Andreas Bartels.

[7] See Carnap 1950, §3.

[8] See Goodman 1983, 66.

is thought to be fallacious, confirmation theory should not mirror it. This is illustrated by many cases of misguided or misinterpreted statistical reasoning. Confirmation theory should be able to *correct* fallacious inferences in the empirical sciences. On the other hand, there is a descriptive component, too: confirmation theory has to extract its rules from practice and should be able to model historical cases of theory confirmation and to bridge the gap to the practice in the empirical sciences.

1.2 The new riddle of induction

The last section has pointed out, *pace Goodman*, that instead of giving an ultimate justification for inductive inference, the proper problem of induction consists in defining the correct rules of inductive inference. But even if we presuppose inductive principles of inference, most accounts of confirmation are underdetermined in an embarrassing way, as pointed out by Goodman (1983). Assume that all emeralds that have been observed so far are green. By all means, this seems to entitle the inductive inference to the conclusion that emeralds to be observed in the future will be green. Now consider the predicate 'grue'. It applies to all objects that were examined in the past (= before t_0) in case they are green and to all other hitherto unexamined objects in case they are blue:

$$\mathrm{Grue}(x) \equiv \begin{cases} \mathrm{Green}(x) & x \text{ examined before } t_0 \\ \mathrm{Blue}(x) & \text{otherwise.} \end{cases}$$

It is very plausible to adopt the view that universal hypotheses like 'all emeralds are green' are confirmed by their instances (as does Hempel [1945] 1965), i.e. the observation of a green emerald, and almost all accounts of confirmation agree on that. However, this leads into problems. Surely, the hypothesis 'all emeralds are green' is confirmed when all emeralds examined so far have turned out to be green. But our *epistemic intuitions* resist to say that 'all emeralds are grue' is confirmed by the observations of green emeralds in the past. Instead, we would like to say that only the first, 'natural' hypothesis that all emeralds are green is confirmed by the past observations. Unfortunately, almost all accounts of confirmation recognize the observations of green emeralds in the past (=before t_0) as instances of the 'green' hypothesis as well as of the 'grue' hypothesis. Goodman's objection to existing accounts of confirmation is not a *logical*, but an *epistemological* one – the principle of confirmation by instances which is prima facie plausible conflicts with our epistemic intuitions.[9] Goodman's critique points out that the principle of induction, allowing for a projection from the past to the future, does not specify which *kinds of projections* are admissible. Both the 'green' and the 'grue' hypothesis make use of an inductive inference, but there seems to be a difference: Projecting the 'green' hypothesis seems to be sound, in contrast to the 'grue' hypothesis. I would like to anticipate that this objection can be applied in a slight variation to other accounts of confirmation, too. So the problem extends beyond the specific Hempelian account of confirmation that was the historical target of Goodman's criticism.

The natural reply to Goodman's challenge consists in saying that only *lawlike* hypotheses – those which use only 'natural' and no gerrymandered predicates – are confirmed in practice. In fact, the 'green' hypothesis, but not the 'grue' hypothesis seems to correspond to a natural law. The 'grue' hypothesis seems to be purely *accidental*. In particular, the predicate 'grue' is gerrymandered from 'green' and 'blue' and involves reference to a particular point of time. It is composed of various 'timeslices' (green before t_0, blue afterwards). This seems to be an indicator for accidental hypotheses – lawlike hypotheses should not be 'gerrymandered' and be uniform in space and time. Therefore,'the 'grue' predicate does not seem to be admissible for genuine confirmation. Goodman responds, however, that we could define a second gerrymandered predicate 'bleen' which is dual to the predicate 'grue':

$$\mathrm{Bleen}(x) \equiv \begin{cases} \mathrm{Blue}(x) & x \text{ examined before } t_0 \\ \mathrm{Green}(x) & \text{otherwise.} \end{cases}$$

Then Goodman notes that

> "true enough, if we start with 'blue' and 'green', then 'grue' and 'bleen' will be explained in terms of 'blue' and 'green' and a temporal term. [...] But equally truly, 'green', for example, applies to emeralds examined before $t_{[0]}$ just in case they are grue, and to other emeralds just in case they are bleen. Thus qualitativeness [not being gerrymandered, J.S.] is an entirely relative matter and does not by itself establish any dichotomy of predicates."[10]

[9] See Fitelson 2008.

[10] Goodman 1983, 79-80.

In other words, the reference to temporal elements in the predicates 'grue' and 'bleen' is entirely relative to an antecedent choice of primitive predicates; those gerrymandered predicates could, by stipulation, be regarded as primitive so that the predicates 'green' and 'blue' would count as gerrymandered. The situation is completely symmetrical – there is no logical reason to claim that the 'green'/'blue' predicates are natural whereas the 'grue'/'bleen' predicates are gerrymandered. On logical grounds, we cannot deny confirmation to 'all emeralds are grue' and to affirm it for 'all emeralds are green' since all formal criteria of lawlikeness are relative to a choice of primitive predicates.

This 'new riddle of induction', as Goodman calls it, can be generalized. For any set of observations that are instances of a certain predicate, there are infinitely many interdefinable predicates that stand in the same relation to the available evidence. For instance, the observation that all emeralds examined so far are green does not only support the hypothesis that all emeralds are green, but also the hypotheses that all emeralds are grue, gred (where 'gred' is similarly defined as 'grue' before), etc. In other words, a nearly arbitrary set of hypotheses is, on nearly all formal accounts of confirmation, equally confirmed. This sounds plainly absurd and sheds a new light on the relationship between the induction problem and confirmation theory. Recall that we aimed at a logical separation of confirmation theory and replies to the induction problem. Now it seems that Goodman's new riddle is almost a reductio of very basal intuitions in confirmation theory: Arbitrary theories are confirmed by arbitrary evidence. Is it at all possible to distinguish some inductive inference amongst an infinity of competing inferences? In his treatment of the induction problem, Hume tried to comfort the sceptical doubts by noting that only those predictions conforming to past regularities, to our *customs* and *habits*, would give rise to valid inductive inferences. But, or so Goodman objects,

> "Hume overlooks the fact that some regularities do and some do not establish such habits; that predictions based on some regularities are valid while predictions based on other regularities are not. [...] To say that valid predictions are those based on past regularities, without being able to say which regularities, is thus quite pointless."[11]

In other words, the Humean problem of induction (is it justified to project

[11] Goodman 1983, 82.

the past into the future?) has been replaced by a *problem of confirmation*: to which method of projection, to which selection of 'natural' predicates should the formal apparatus of confirmation theory be applied? Confirmation theory therefore needs an antecedent selection of 'natural' predicates whose instances can be projected and which can be (dis)confirmed in the ordinary way. This residual problem of distinguishing lawlike from accidental hypotheses is leaved to another discipline – we are solely interested in the formal characterization of confirmation on the grounds of a set of primitive natural predicates.[12] The assumptions on which confirmation theory stands are actually much stronger than the minimal inductive assumption that nature is uniform in time and that past regularities can be projected into the future. It has to be presupposed in which way nature is uniform in time and which regularities are projected. The usual tool of accomplishing that is done by choosing and shaping a logical language – for instance, a lot of confirmation theory takes place in the framework of first order logic and the predicates of that language are simply assumed to be projectible. This does not circumvent the new riddle of induction, but it puts the problem into parentheses: *If* the predicates are projectible, *then* the so-and-so account describes the rules of valid inductive inference, opening a way to model real cases of scientific confirmation. A formal confirmation theory acknowledges the open problems, but it does not abandon the project of finding the rules of valid inductive inference.

1.3 The task of confirmation theory

Scientists frequently disagree whether an empirical finding really confirms a theoretical hypothesis. Common scientific sense may not be able to settle the question, first because the case under scrutiny might be very complicated and second, because people might have different ideas of common sense in a specific case. Formal criteria of confirmation would help to settle the discussion, and once again, it is helpful to consider the analogy to deductive logic. Whether a deductive inference is valid can be decided by applying formal criteria and mechanical procedures – every valid formula can be deduced from the logical axioms (that is Gödel's famous completeness theorem). Hence,

[12] An attempt to determine the set of projectible predicates is made in the fourth chapter of Goodman 1983. By the way, this problem is characteristic of inductive logic and confirmation theory and has no counterpart in deductive logic.

in case there is a disagreement about the validity of a deductive inference, the formal tools can help us to settle the question. In the same way that the validity of a deductive inference can be checked using formal tools (deductions), it is desirable to have formal tools which examine the validity of an inductive inference. Sometimes this project is deemed futile because scientists do not always make their criteria of confirmation explicit. But that objection conflates a logical with a psychological point (Hempel [1945] 1965, 9-10) – analogous to the frequently encountered conflation between context of justification and context of discovery in philosophy of science. Confirmation theory and inductive logic aim at a rational reconstruction of inductive practice that is not only descriptively adequate, but also formally fruitful and able to correct mistakes in science.

We should acknowledge, however, that confirmation is an ambiguous term with two sub-concepts which fall into its domain: *absolute* confirmation and *relative* confirmation. We often say that a certain theory is well confirmed, but we also say that a certain piece of evidence confirms a hypothesis. These two different usages correspond to different meanings of the word 'confirmation'. When we use the former way of speaking – 'theory T is well confirmed' – we say something about a particular theory: T enjoys high confidence and the total available evidence speaks for T and favors it over all serious rivals. To be confirmed or to be well confirmed becomes a property of a *particular hypothesis* or theory. By contrast, the latter use says something about a *relationship* between hypothesis and evidence – it is asked whether a piece of evidence supports or undermines a hypothesis. Relative confirmation means that an empirical finding, a piece of evidence, lends support to a hypothesis or theory. This need, however, not imply that on account of the total available evidence, the theory is well confirmed. Several reasons speak for focussing one's interest on relative confirmation: First, it is plausible that absolute confirmation is secondary to relative confirmation: a theory is absolutely confirmed if and only if it is relatively confirmed to a sufficiently strong degree by the total available evidence. Therefore we have to study relative confirmation if we want to understand when a theory is well-entrenched and well-confirmed. Second, relative confirmation plays a much larger role in scientific practice than absolute confirmation: when we perform an experiment and proceed to the evaluation, we would like to assess the relationship between the observed data and the hypothesis under test and to find out whether and to which degree the data support the hypothesis, regardless of how credible was the hypothesis before. A suitably large number of experiments where relative confirmation takes place can lead us to the belief that a certain hypothesis is absolutely confirmed and highly credible, but the converse does not hold: Absolute confirmation does not play a role in determining whether a piece of evidence confirms a hypothesis. Third and last, the notion of relative confirmation helps us to examine more general issues in philosophy of science that touch the relationship between theory and models on the one hand and data and phenomena on the other hand. Such questions might be the problem of epistemic holism in science or the theory-ladenness of observation. We will get back to all these issues in the course of the book.

Once the relational character of confirmation is clarified, I feel the need to say something about the objects of this relation. At first sight, it sounds plausible to think of confirmation as a *semantic* relation between a first-order sentence on the one side – the scientific hypothesis – and a real-world object on the other side. For instance, a black raven seems to confirm the hypothesis that all ravens are black. On the evidential side, we have objects, and on the theoretical side, we have first-order sentences. But if we pursue the project to assimilate confirmation theory to deductive logic and to find a system of *syntactic* rules for valid inductive inference, a semantical relation is clearly inadequate. As long as the evidence is an external object it cannot figure in a syntactic relation.[13] But framing the evidence into sentences of a formal language gives us access to powerful logical tools, e.g. we would be able to check whether the evidence can be deduced from the hypotheses or whether it is consistent with the hypothesis. Working with real-world objects as evidence would deprive us of all those tools and furthermore, the evidence has to be expressed and communicated in sentences of a natural or formal language. So if we aim at a model of scientific confirmation, at a calculus for truth-conducive (though not truth-guaranteeing) inference, we should establish syntactic relations and treat both sides as formulae of first-order logic. The analogy to deductive logic which successfully works on a syntactic level illuminates the point.

Making confirmation relative to a certain language allows, of course, for some ambiguity: Different people will put the content of a 'real world event' into different linguistic frameworks. So, dependent on the language which

[13] See Hempel [1945] 1965, 21-22.

is employed, the same event might lead to different 'observations', different formulations of the evidence and finally different results. But this objection does not need to trouble us: History of science (e.g. the tenth chapter in Kuhn 1962) teaches us that scientists of different ages who perform the same experiment have seen different things because they were placed in different *paradigms*, implying a different vocabulary for describing their observations. For example, the Aristotelians described the motion of a pendulum as a constrained fall whereas Galilei was able to see a periodic oscillation. So language dependency and paradigm dependency are quite natural things and any account of confirmation can only work within a paradigm and the associated scientific language. – For roughly the same reasons, I do not conceive the relata of the confirmation relation as abstract objects like propositions in a Platonic heaven. First, this would deprive us of the analysis tools that are available for sentences and second, it would beset the entire projectwith a metaphysical burden (what are propositions? what is their ontological and epistemic status?). Therefore I follow the main tradition in confirmation theory – the confirmation relation holds between *sentences of a formal language* $\mathcal{L}$. This relation is a purely structural relationship between evidence and hypothesis in the sense that the confirmation relation holds regardless of the meaning which we assign to the language parameters. The confirmation relation is then characterized by means of relations between well-formed formulae of $\mathcal{L}$, independent of the chosen structure. In this respect, confirmation theory is a logic of confirmation, analogous to deductive logic where valid inferences are precisely those which hold in any structure of $\mathcal{L}$.

We have seen that the predicate of absolute confirmation applies to a hypothesis alone whereas the predicate of relative confirmation says something about the *relation* between hypothesis and evidence. So the predicate of relative confirmation (which is the predicate we are interested in) has one more place than the predicate of absolute confirmation. But that is not the end of the story. In determining the confirmation relation between hypothesis and evidence we often make tacit reference to a corpus of background knowledge. Take Kepler's second law of planetary motion – a line joining a planet and the sun strikes equal equal areas in a fixed interval of time, irrespective of the planet's position in his orbit. In particular, the closer planets get to the sun in their period, the faster they travel. It is possible to derive Kepler's second law from the Newtonian theory of gravitation, but at the time Kepler proposed his laws, that theory was not yet invented and Kepler had to draw

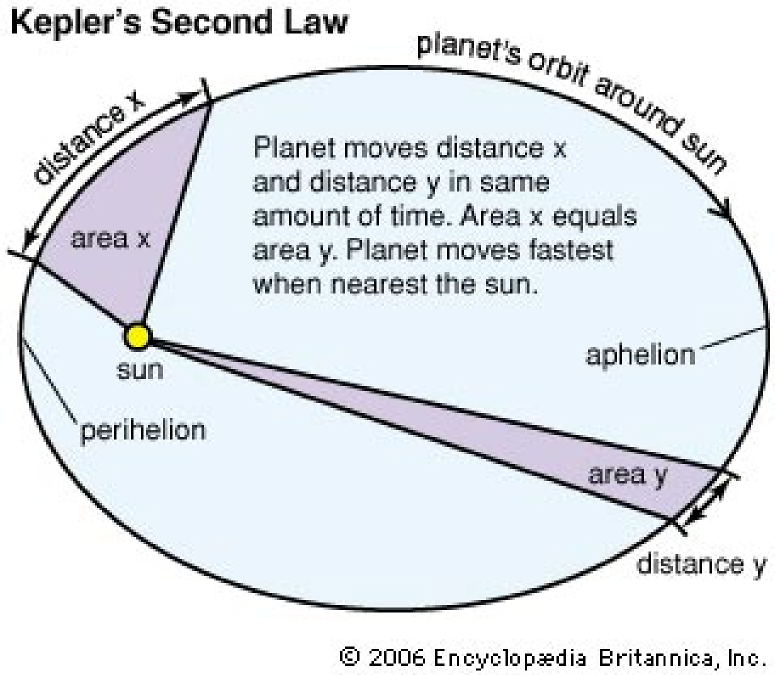
Kepler's Second Law
planet's orbit around sun
distance x
area x
Planet moves distance x and distance y in same amount of time. Area x equals area y. Planet moves fastest when nearest the sun.
sun
perihelion
aphelion
area y
distance y
© 2006 Encyclopædia Britannica, Inc.

Figure 1.1: Kepler's second law in a graphical representation. Source: Encyclopædia Britannica (www.eb.com).

on the available observation data. He had to confirm it by observation, using lots of observation data about planetary motion. But of course, Kepler was not able to directly see the second law – when we look at the sky we do not literally see that planets travel in ellipses and that the sweeped areas are equal. Thereby he had to draw on a lot of auxiliary assumptions – that the data had ben gained from reliable instruments, that the Copernican model of the solar system was basically correct and that planets travel in ellipses around the sun (Kepler's first law), that the methods to infer the distance of the planets from the sun were correct, and so on. Those auxiliary hypotheses connected his abstract hypothesis to the available data.

Background knowledge and auxiliary assumptions enable us to judge whether a piece of evidence confirms a hypothesis or not. Had the background assumptions been different, there might have been no (dis)confirmation relation between our observations and the theory. For instance, if we did not know a method to determine the distance of the planets to the sun, we would not be able to confirm or disconfirm Kepler's second law by means of the observational data that were available to seventeenth-century astronomers. So confirmation is *relative* to the kind of assumptions we make and to the background with regard to which we evaluate this predicate. These assumptions typically contain initial conditions, claims about the measurement apparatus and auxiliary laws which gap the bridge between theoretical hypotheses and observable consequences. Another equally famous example is the confirmation of Einstein's hypothesis that light is bent by massive bodies, a corollary of the General Theory of Relativity. This claim entailed that passing starlight would be bent by the sun to an angle of 43 seconds per arc, about the double of what Newtonian physics predicted. Such an effect becomes, of course, visible only during an eclipse. Famously, Eddington successfully checked the prediction of GRT in 1919, and by comparing his photographs to pictures of the relevant sky region that were taken at night, he was able to spot a difference between the pictures and to show that Einstein was right. But again, Eddington required a lot of auxiliary assumptions, e.g. about the effect size of atmospheric starlight aberration.

Of course, there are cases of confirmation where almost no background assumptions are required and the hypothesis is close to the observations. Statistical hypothesis often combine various forms of background assumptions into a single claim. But in general, science builds on a lot of auxiliary assumptions. They are the less dispensable the more theoretical a hypothesis under test becomes.[14] So we are well advised to consider the background assumptions as an integral part of the confirmation relation. Moreover we will see in the next chapters that a two-place predicate of confirmation has serious deficits and cannot deliver an adequate theory of confirmation. Without going into the technical details, we can state at this point that researchers typically make a distinction between hypotheses *under test* and hypotheses *in use* and that a two-place predicate of confirmation neglects that distinction, thus failing to be descriptively adequate. In order to keep the presentation clear and readable, I will not always explicitly mention the background knowledge in the book. Indeed, conditioning on background knowledge is an important tacit assumption in confirmation theory.

Thus, the topic of the book is the relationship between theory and evidence, between models and data, relative to certain background assumptions or certain background knowledge. This approach does not only enable a more fruitful and penetrating analysis of the concept of confirmation, as argued above, it also opens a natural way to integrate statistical evidence and appli-

[14] See Duhem 1914, in particular p. 281.

cations of statistical evidence in various problem areas of philosophy. Since statistics and statistical methods are pervasive in the empirical sciences, no analysis of confirmation that is interested in implications for real science can neglect these issues. By means of the probabilistic framework, we also obtain a quantification of the degree of support. When discussing statistical evidence we will also notice the distinction between *subjective* and *objective* accounts of confirmation – accounts which allow for distinctly subjective elements in the confirmation of a scientific hypothesis and those who do not. Scientific inference is supposed to be as objective and intersubjectively binding as possible, and this normative force of a scientific conclusion should not be put in jeopardy by making the concept of confirmation too subjective. But on the other hands, it is questionable whether an 'objective' theory of inductive inference can ever be achieved.

All those questions will be discussed in the latter chapters of the book. I would like to begin with the attempt to straightforwardly explicate the confirmation predicate in a framework where many scientific theories can be embedded: first order predicate logic. This is the subject matter of *qualitative confirmation* – accounts of confirmation that do not come with a numerical quantification of the degree of support lent by the evidence. They are putatively objective theories of confirmation and the most traditional tool of tackling the problem of scientific confirmation.

1.4 Summary

The problem of induction – how to justify the projection from past to future regularities – has been a vexing issue over a long time, but a satisfactory resolution is still to be found. Two hundred years after David Hume formulated the problem in a pressing way, philosophers have got used to living with it and not spending too much worries on its resolution. Confirmation theory can, or so the received view argues, nevertheless extract rules from our inductive practice and in turn measure our practice against these rules. However, Goodman's 'new riddle of induction' directly affects confirmation theory, too: the main question is not *whether* to project past regularities to future expectations but *which* kind of regularities should be projected. Goodman argues that confirmation theory alone cannot succeed in distinguishing valid projections from invalid one, as famously illustrated by the 'grue' example. Our epistemic intuitions urge us to project 'lawlike' regularities (like 'green') and to forbid the projection of 'accidental' regularities (like 'grue'). Goodman's argument shows that formal accounts of confirmation which are supposed to distinguish valid from invalid inductive inferences cannot solve this problem. Hence confirmation theory has to build on assumptions which predicates are projectible and which are not – a concession that is actually less dramatic than it sounds.

Confirmation theory tries to establish formal criteria for confirmation of a scientific theory and to model the inductive practice of scientists. The project of finding a formal explication of confirmation and rules for valid inductive inference is similar to the project of finding deductive rules that characterize a valid (deductive) inference. Here, we have to distinguish the relative concept – 'What is the relation between a piece of evidence and a hypothesis?' – and the absolute concept of confirmation – 'Is a hypothesis credible and well-entrenched?'. We have argued that the relative concept of confirmation is the more fundamental concept and the more interesting object of study. Moreover, due to the enormous role of auxiliary hypotheses in science, we add a third element to the confirmation relation – background assumptions. These preliminaries pave the way for the second chapter where qualitative, structural accounts of confirmation are discussed.

Chapter 2

Qualitative Confirmation

2.1 Introduction

In the last two decades, qualitative accounts of confirmation have largely been superseded by probabilistic accounts, in particular Bayesian ones. While probabilities certainly provide a powerful framework for inductive reasoning, this does not imply that qualitative reasoning has become superfluous. In a lot of empirical sciences probabilistic reasoning still plays a minor role, if at all. Qualitative arguments are thus central for the confirmation of scientific hypotheses, and their relativity to a set of primitive predicates of a language does not diminish their significance. They allow us to reconstruct cases of confirmation in science and develop normative constraints for theory confirmation. Indeed, the most prominent cases of theory confirmation and replacement are situated in a qualitative framework, e.g. the confirmation of Kepler's laws of planetary motion or Darwin's theory of evolution. In order to have a sensible model for such cases, an account of qualitative confirmation is indispensable – introducing subjective probabilities would simply misrepresent the problem. Replacing qualitative by quantitative accounts in the entire domain of science was succinctly criticized by Clark Glymour:

> "The bearing of evidence on theory is thought to be established by probabilistic connections, and confirmation and methodology are to be explicated in probabilistic terms. [...] Such arguments slide over much of the structure of scientific arguments that we find in fact, and impose instead a probabilistic superstructure. In providing an admirable, general account of confirmation and rational belief, probabilistic theorists are obliged largely to ignore,

> for example, the intricacies of Newton's argument for universal gravitation and of many other scientific arguments of major interest."[1]

Of course, quantitative (e.g. probabilistic) theories of confirmation are always qualitative theories, too, so Glymour's objection seems to be off the mark.[2] Nonetheless, I believe that there are two important points in favor of purely qualitative accounts: first, as pointed out by Glymour, a representation of confirmatory arguments in the history of science by means of degrees of belief is often inadequate – just because the arguments actually went another way. This is especially plausible for non-stochastic theories as Newtonian mechanics. Second, probabilistic confirmation theories often allow different verdicts on the strength of a confirmatory argument, but in many cases, the confirmatory power of the evidence seems to be beyond reasonable doubt and to leave little room for subjectively based disagreement. Purely qualitative theories of confirmation are, as we will see, more objective than quantitative, probabilistic ones based on degrees of belief. They are much better able to capture *structural relations* between theory and evidence that are essential to scientific confirmation:

> "[...] that relation [of evidential relevance] depends somehow in *structural, objective features connecting statements of evidence and statements of theory.* [...] There must be relations between evidence and hypotheses that are important to scientific argument and to confirmation but to which the Bayesian scheme has not yet penetrated."[3]

This gives a rationale for focussing on qualitative accounts, so much the more as many empirical (and in particular, physical) theories are formulated inside the framework of first-order logic. Thus, if we want to capture structural relations of evidential support, it is natural to take first-order logic as a framework for non-probabilistic, qualitative confirmation theory, too. In sharp contrast to probabilistic theories of confirmation, qualitative theories of confirmation do not try to measure the degree of confirmation. Instead they set up conditions for the question when something counts as evidence

[1] Glymour 1980a, 5.

[2] Rainer Stuhlmann-Laiesz urged me to clarify that point.

[3] Glymour 1980a, 93, my emphasis.

for a hypothesis. Such qualitative accounts are often closer to scientific practice than the more expressive and fine-grained probabilistic account which in many cases merely impose a 'probabilistic superstructure'. Of course, the power of such accounts is restricted because nowadays, many scientific hypotheses have statistical character so that a comprehensive confirmation theory requires the more expressive language of quantitative confirmation and probability. Still, it is an attractive and worthwhile project to give an 'objective' account of confirmation, without recourse to subjective probabilities. Apart from that, study of qualitative confirmation reveals the role of deductive relations in scientific reasoning and illuminates typical features of confirmation. Learning how the concept of confirmation works gives helpful hints for a quantitative analysis, too (which is conducted in the later chapters of this book). All these facts encourage philosophers of science not to give up qualitative confirmation and to keep the field alive.

The outline of the chapter is thus. First, I present the two major approaches in qualitative confirmation theory – confirmation by instances, as captured in the satisfaction criterion and the hypothetico-deductive tradition. Thereby I show that both approaches are subject to severe difficulties, some of technical and some of a more principled nature. Then I discuss some attempts to solve the problems before finally coming to my own suggestion which aims at a reconciliation of falsificationist principles with confirmation theory. This is accomplished in a refined account of hypothetico-deductive confirmation that incorporates falsificationist principles to a higher degree than its predecessors.

2.2 Hempel and the raven paradox

A very natural idea to explicate the concept of qualitative confirmation consists in the idea that hypotheses are confirmed by finding their *instances*. This approach was first suggested by Jean Nicod [1923] (1970) and has been influential up to modern times (e.g. Glymour 1980a). The idea is thus: When you have a hypothesis that all X's are Y's, this hypothesis is confirmed by the observation of a X that is also a Y. For example, if we want to confirm the hypothesis that all ravens are black, only the observation of a black raven seems to confirm that hypothesis. In other words, such universal conditionals as 'all ravens are black' or, more formally, $\forall x : Rx \rightarrow Bx$ are confirmed by any observation of the form $Ra.Ba$ and by nothing else. Such an account seems to correspond precisely to our intuitions about confirmation.

> **Nicod Condition (NC):** For a hypothesis of the form $H = \forall x : Rx \rightarrow Bx$ and any individual constant a, an observation report of the form $Ra.Ba$ confirms H (relative to empty or irrelevant background knowledge K).

However, this account runs into serious trouble. To see why, I would like to motivate the

> **Equivalence Condition (EC):** If H and H' are logically equivalent sentences then E confirms H relative to K if and only if E confirms H' relative to K.[4]

We have already said that scientific hypothesis are usually framed in the logical vocabulary of first-order logic (or a reduct thereof). Furthermore, they are often stated in different, but logically equivalent forms, e.g. the mathematical property of compactness can be stated using a topological or an equivalent analytical formulation. The idea of the equivalence condition is that 'saying the same with different words' does not make a difference with regard to relations of confirmation and support: Hypotheses which express the same content with different words are equally supported and undermined, independent of the chosen formulation. For instance, if we assert that set S is compact, the amount of (dis)confirmation another sentence lends to this assertion is independent of whether we have analytical or topological compactness in mind. To see this in more detail, note that for deductive relations (e.g. whether A logically implies B), the Equivalence Condition holds by definition: If A logically implies B, A also implies any B' that is logically equivalent to B. An account of confirmation should contain relations of deduction and entailment as special cases: If an observation entailed the negation of a hypothesis, in other words, if the hypothesis were *falsified* by actual evidence, this would equally speak against all equivalent versions and formulations of that hypothesis. Deduction and logical entailment do not make a difference between equivalent sentences, and such logical relationships between hypothesis and evidence are clearly relevant for confirmation

[4] This condition can naturally be extended to a condition for evidence and background knowledge, asserting that the confirmation relation is invariant under replacing evidence/background knowledge statements by logically equivalent statements. For the purposes of discussing the raven paradox, this is, however, not necessary.

in science. For instance, it is of major interest whether Kepler's laws are a logical consequence of Newton's laws of motion and gravitation. Since the Equivalence Condition holds for deductive relations and confirmation has to build on them, we are well advised to demand the Equivalence Condition for inductive relations, too. It would be very strange if choosing a different formulation of a natural law invalidated the confirmation which empirical evidence lent to the law. If the equivalence condition did not hold, the degree of support would depend on the specific formulation of the law which would counter all efforts for introducing logical and mathematical methods into science, thereby making it more rigorous and finally more successful.

Hence, the equivalence condition should be accepted without contention. This leads, however, to a problem for Nicod's suggestion that hypotheses are confirmed by observing their instances: Assume that we observe $E = \neg Ra.\neg Ba$ (e.g. imagine that E is a piece of white chalk). Obviously, E confirms the hypothesis $H_2 = \forall x : \neg Bx \rightarrow \neg Rx$ – things that are not black are no ravens either (by the Nicod condition). Furthermore H_2 is logically equivalent to its contrapositive $H_1 = \forall x : Rx \rightarrow Bx$ – the original raven hypothesis. Hence, by the equivalence condition, E confirms H_1, too. This conflicts with our intuition H_1 – the hypothesis that all ravens are black – should not be confirmed by observing a piece of white chalk (which has the properties of neither being a raven nor being black). Hence, we have three individually plausible, but incompatible claims at least one of which has to be rejected:

1. **Nicod Condition (NC):** For a hypothesis of the form $H = \forall x : Rx \rightarrow Bx$ and any individual constant a, an observation report of the form $Ra.Ba$ confirms H.

2. **Equivalence Condition (EC):** If H and H' are logically equivalent sentences then E confirms H relative to K if and only if E confirms H' relative to K.

3. **Confirmation Intuition (CI):** A Hypothesis of the form $H = \forall x : Rx \rightarrow Bx$ is not confirmed by an observation report of the form $\neg Ra.\neg Ba$.

The main conflict in this sets of claims consists in the fact that (EC) merely considers the *logical form* of scientific hypothesis whereas (NC) and (CI) implicitly assume that there is an 'intended domain' of a scientific hypothesis.

In particular, only ravens are allowed to confirm or disconfirm the hypothesis that all ravens are black.

While we will discuss the raven paradox in greater detail later it is interesting to note in the first place that in his [1945] 1965, Hempel argues against (CI). We should learn to live with the paradox and not see anything paradoxical in the fact that something that is neither a raven nor black confirms the hypothesis that all ravens are black. His argument can be paraphrased thus:[5] Assume that we observe a grey, formerly unknown bird that is in all relevant external aspects very similar to a raven. That observation puts the raven hypothesis to jeopardy. It might thus be the case that we have seen a non-black raven and have thus falsified our hypothesis. But a complex genetic analysis reveals that the bird does not belong to the kind of ravens – indeed, it is more related to crows than to ravens. Hence, it sounds logical to say that due to the results of the genetic analysis, the observation of the grey crow corroborates the raven hypothesis – the raven hypothesis has survived a possible falsification. In other words, a potential counterexample has been eliminated. Thus there is no paradox in saying that an observation report of the form $\neg Ra.\neg Ba$ confirms H – in the sense that a satisfies the constraint given by H that nothing can be both a raven and have a color different from black. It might now be objected that the observation of a black raven seems to lend *stronger* support to the raven hypothesis than the observation of a grey crow-like bird. But this is a problem for a quantitative analysis of the raven paradox and not for a qualitative one. By not being a counterexample to H the observation of the grey crow-like bird supports H, and this is particularly salient when $\neg Ba$ is learnt before $\neg Ra$. Hence (CI) is – at least from the point of view of a qualitative theory of confirmation – plainly false. The paradox vanishes since one of the three premises has been discarded.

Now we have to look for a full account of confirmation that respects (EC), (NC) and the failure of (CI). To this end, Hempel suggests a further criterion of adequacy for confirmation: When we confirm a hypothesis H, the confirmation transmits to any hypothesis that is logically *weaker* than H, i.e. to any hypothesis that is entailed by H. That is again motivated by the analogy to deductive logic where, if $\Gamma \models \phi$, any logical consequence of ϕ is also entailed by Γ.

[5] Hempel [1945] (1965) makes the argument for quite a different example ('all sodium salts burn yellow') but I would like to stick to the original raven example because I do not want to confuse the reader.

> *Consequence Condition (CC)*: If E confirms H relative to K and H logically entails H' ($H \models H'$) then E confirms H' relative to K, too.

For instance, an observation that confirms the heliocentric model of the solar system (as Galilei's discovery of the Jupiter moons) also confirms the special corollary that Earth revolves around the sun. Or observation data that confirm the wave nature of light implicitly confirm that light exhibits diffraction patterns. In other words, an observation that makes us confident in a strong and comprehensive theory makes us also more confident in its parts. We will discuss that intuition later, but for the moment, we take it as granted.

Building on these and some other criteria for confirmation, Hempel develops a full account of qualitative confirmation, the satisfaction criterion. The idea is thus: Deductive entailment between evidence and hypothesis is certainly too strong as a criterion of confirmation, but we may wish to say that the evidence entails a restricted part of the hypothesis – namely the part which that observation is able to verify. For example, if a confirming observation report says something about the singular terms a, b and c, the claims a hypothesis makes about a, b and c have to be *satisfied* by the evidence. From such an observation report we could conclude that the hypothesis is true of the class of objects that occur in E. That is all we can demand of an confirming observation report, or so Hempel argues. In other words, we gain *instances* of a hypothesis from the evidence, and such instances confirm the hypothesis. To make this informal idea more precise, we have to introduce some definitions (taken from Gemes 2006a):

Definition 2.1 *An atomic well-formed formula (wff) β is* relevant to a wff *α if and only if there is some model M of α such that: if M' differs from M only in the value β is assigned, M' is not a model of α.*

So intuitively, β is relevant for α if at least in one model of α the truth value of β cannot be changed without making α false. Now we can define the domain of a wff:

Definition 2.2 *The* domain of a well-formed formula *α, denoted by $dom(\alpha)$, is the set of singular terms which occur in the atomic (!) well-formed formulas (wffs) of L that are relevant for α.*

For example, the domain of $Fa.Fb$ is $\{a, b\}$ whereas the domain of $Fa.Ga$ is $\{a\}$ and the domain of $\forall x : Fx$ are all singular terms of the logical language. In other words, quantifiers are treated substitutionally. The domain of a formula is thus the set of singular terms about which something is asserted. Those singular terms are said to occur essentially in the formula:

Definition 2.3 *A singular term a* occurs essentially *in a formula β if and only if a is in the domain of β.*

So, i.e. a occurs essentially in $Fa.Fb$, but not in $(Fa \vee \neg Fa).Fb$. Now, we are interested in the development of a formula for the domain of a certain formula.

Definition 2.4 *The* development *of a formula H for a formula E, $H_{|E}$, is the* restriction *of H to E, i.e. the restriction of H to the domain of E or all singular terms that occur essentially in E. In particular, the restriction of the formula $\forall x : Fx$ is is satisfied if and only if $\forall x : (x \in dom(E) \rightarrow Fx)$.*[6]

In other words, the *development* of a hypothesis for a set of singular terms is the restriction of H to that set. For instance, $(\forall x : Fx)_{|\{a,b\}}$ becomes $Fa.Fb$. Now we have the technical prerequisites for understanding Hempel's satisfaction criterion:

Definition 2.5 *(Satisfaction criterion) A piece of evidence E* directly Hempel-confirms *a hypothesis H relative to background assumptions K if and only if $E.K$ entails the development of H to the domain of E. In other words, $E.K \models H_{|dom(E)}$.*

Definition 2.6 *A piece of evidence E* Hempel-confirms *a hypothesis H relative to background assumptions K if and only if H is entailed by a set of sentences Γ so that for all sentences $\phi \in \Gamma$, ϕ is directly Hempel-confirmed by E relative to K.*

So, for example, Fa (directly) Hempel-confirms the hypothesis $\forall x : Fx$ and $Ra.Ba$ and $\neg Ra.\neg Ba$ both confirm the 'raven hypothesis' $H = \forall x : Rx \rightarrow Bx$, in agreement with Hempel's rendering of the raven paradox. Obviously, every piece of evidence that directly-Hempel confirms a hypothesis

[6] The development of a formula can be defined precisely by a recursive definition, see Hempel 1943. For our purposes, the informal version is sufficient.

also Hempel-confirms it, but not vice versa. Most intuitively clear cases of confirmation are successfully reconstructed in Hempel's account.

It is easy to see that any sentence that follows from a Hempel-confirmed sentence is Hempel-confirmed, too.[7] Hence, Hempel's confirmation criterion satisfies the Consequence Condition. Similarly, it satisfies the Equivalence Condition because it builds on relations of logical consequence which are invariant under equivalent transformation. However, there are some serious drawbacks of Hempel's classical suggestion.[8]

First, some hypotheses do not have finite developments and are therefore not confirmable. Take the hypothesis

$$H_2 = (\forall x : \neg Gxx).(\forall x : \exists y : Gxy).(\forall x, y, z : Gxy.Gyz \rightarrow Gxz)$$

which asserts that G is a serial, irreflexive and transitive two-place relation. These properties entail that H_2 is not satisfiable in any finite structure and thus not Hempel-confirmable by a finite number of observations. But certainly, H_2 is not meaningless – you might interpret G as the 'greater than' relation and then, the natural numbers with their ordinary ordinal structure are a model of H_2: H_2 would assert the 'greater than' relation is transitive, irreflexive and that for any natural number, there is another natural number which is greater than it.

Second, consider c, an individual constant of our predicate language, and the hypotheses $H_3 = \forall x : Ix$ and $H_4 = \forall x \neq c : \neg Lx$. Take the set of all planets of the solar system as the universe of our intended structure and let the individual constant c refer to Planet Earth. Then H_3 might be interpreted as the claim that iron exists on all planets and H_4 as the claim that no life exists on other planets. Both are meaningful hypotheses open to empirical investigation. Now, the observation report $E = Ic$ (there is iron on Earth) directly Hempel-confirms $H_3.H_4$ (there is iron on all planets and life does not exist o other planets) relative to empty background knowledge.[9] While this may be acceptable, it also follows that H_4 is Hempel-confirmed by $E = Ic$, due to the consequence condition. This is utterly strange since the actual observation (there is iron on Earth) is completely independent of the

[7] More precisely, assume that $H \models H'$ where H is Hempel-confirmed by E relative to K. Then there is a set Γ so that any element of Γ is directly Hempel-confirmed by E (relative to K) and that $\Gamma \models H$. Since by assumption $H \models H'$, it follows that $\Gamma \models H'$, too. Thus H' is Hempel-confirmed.

[8] See also Earman and Glymour 1992.

[9] The development of $H_3.H_4$ with regard to c is Ic.

hypothesis at stake (no life exists on other planets). Clearly, this conclusion goes beyond what the available evidence entitles us to infer and is sheer nonsense.

These formal problems may be mitigated in a more refined formulation of Hempel-confirmation, but there are more fundamental problems, too. They are in a similar vein connected to the fact that Hempel-confirmation satisfies the Consequence Condition. When a hypothesis H is Hempel-confirmed by a piece of evidence E (relative to K), any arbitrary disjunction X can be tacked to H while leaving the confirmation relation intact. For example, the hypothesis that all ravens are black *or* all doves are white is Hempel-confirmed by the observation of a black raven, although it is not clear in how far that observation is *relevant* for the hypothesis that all doves are white. Even worse, the same observation also confirms the hypothesis that all ravens are black or *no* doves are white. The tacked disjunction is completely arbitrary. Evidential relevance for the hypothesis gets lost, but a good account of confirmation should take care of these relations.

Finally, consider the following case: A single card is drawn from a standard deck. We do not know which card it is. We consider, however, the two hypothesis that the card is the ace of diamonds (H_5) and that the card is a red card at all (H_6). Now, the person who draws the card tells us that the card is a diamond and either an ace or a king. Obviously, the hypothesis H_6 is conclusively Hempel-confirmed by this observation report. But what about H_5? We are now much more confident that H_5 is true because the evidence supports the hypothesis that the card is an ace of diamonds over the hypothesis that the card is no ace of diamonds, in the usual *relative* sense of confirmation. However, the observation does not Hempel-confirm the hypothesis that the card is an ace of diamonds. This is so because not all assertions H_5 makes about this particular card – that it is an ace and a diamond – are satisfied by the observation report. This behavior of Hempel-confirmation is somewhat strange and stands in contrast to the most popular quantitative account of confirmation, the Bayesian account. This toy example has analogues in science, too: it is not possible to confirm Kepler's laws in total by confirming only one of its three components. Any confirming observation report would have to entail each of Kepler's laws (with regard to the planet that is observed). This is at least strange because we often do not have the opportunity to check each of the predictions of a theory. Or an observation of the diffraction pattern of light would not confirm the hypothesis that light is an electromagnetic wave because waves have more characteristic properties than just the diffraction pattern – properties that were not shown in that particular observation. This is all very strange. We would like to be able to partially confirm a general hypothesis by successfully checking a particular prediction. Hence, Hempel's satisfaction criterion is not only liable to severe technical objections, but also fails to reconstruct an important line of thought in scientific observation and experimentation.

We thus notice that these counterexamples do not only illuminate technical problems of Hempel's account, but also that the Consequence Condition leads us into big trouble. But why did it seem to be so plausible at first sight? I believe that the missing distinction between the absolute and the relative concept of confirmation is the culprit. The Consequence Condition is plausible whenever *absolute confirmation* is examined. When a strong, comprehensive theory is strongly endorsed – in the sense of 'strongly endorsed' or 'empirically supported beyond all reasonable doubt' – any part of this theory is also absolutely confirmed. Obviously, the less risky a conjecture is, the more confidence can we put in it, and any proper part of a theory is logically weaker and thus less risky than the entire theory. Therefore the Consequence Condition makes perfect sense when it comes to endorsement and absolute confirmation. It is, however, questionable whether the Consequence Condition is also a sensible condition with regard to *relative confirmation*, as the above examples make clear. Here, we are interested in an account of relative confirmation where a theory can also gain support from examining some of its parts. Thus, it is natural to drop the consequence condition and to abandon Hempel's proposal, too. In the next section, we will review the main alternative to Hempel's satisfaction condition, the venerable hypothetico-deductive approach to confirmation.

Finally, Hempel's proposal even fails to resolve the problem which motivated the entire account: the raven paradox. Take again the hypothesis H that all ravens are black. Compare two possible pieces of evidence: In the first case, we take a crow *which we know to be a crow* and notice that it is grey ($E_1 = \neg Ba.\neg Ra$[10], $K_1 = \neg Ra$). This seems to be a fake experiment if evaluated with regard to the raven hypothesis – we knew beforehand that the crow could not disconfirm the raven hypothesis. There was no risk involved

[10] For the confirmation relation it does not make a difference whether we write the evidence as $E_1 = \neg Ba$ or as $E_1 = \neg Ba.\neg Ra$ because K_1 contains $\neg Ra$.

in the experimentation.[11] In the second case we observe a grey crow *which we do not know to be a crow* and realize only by means of a cumbersome genetic analysis that the bird is not a raven, but a crow ($E_2 = \neg Ra.\neg Ba$, $K_2 = \emptyset$). That counts as a sound case of confirmation, as argued above (in agreement with Hempel). Hempel spots the difference as thus: When we are told beforehand that the bird is a crow

> [...] "this has the consequence that the outcome of the [...] color test becomes entirely irrelevant for the confirmation of the hypothesis and thus can yield no new evidence for us."[12]

In other words, the available background knowledge in the two cases makes a crucial difference. Not taking this difference into account is responsible for the fallacious belief (CI) that nothing that is neither a raven nor black can confirm the hypothesis that all ravens are black. (CI) is plausible only if we tacitly introduce the additional *background knowledge* that the test object is no raven. Thus, in the above example, H should be confirmed if we do not know beforehand that the bird under scrutiny is a crow ($K_1 = \emptyset$) and it should not be confirmed if we know beforehand that the bird is a crow ($K_2 = \neg Ra$). In Hempel's own words,

> "If we assume this additional information as given, then, of course, the outcome of the experiment can add no strength to the hypothesis under consideration. But if we are careful to avoid this tacit reference to additional knowledge (which entirely changes the character of the problem) [...] we have to ask: Given some object a (that is neither a raven nor black, but we do not happen to know this, J.S.): does a constitute confirming evidence for the hypothesis? And now [...] it is clear that the answer has to be in the affirmative, and the paradoxes vanish."[13]

However, Hempel is unable to make that difference in his own theory of confirmation. The reason is that his account is *monotone* with regard to the background knowledge, i.e. extending the background knowledge cannot destroy the confirmation relation.[14] Hempel inherits this property from

[11] See Popper 1963.

[12] Hempel [1945] 1965, 19.

[13] Hempel [1945] 1965, 19-20.

[14] This argument was first made in Fitelson and Hawthorne (2006) and Fitelson (2006, 98-99).

deductive logic, because $E.K \models H_{|dom(E)}$ is the crucial condition for direct Hempel-confirmation, and thus also for Hempel-confirmation. Evidently, logical entailment is preserved under adding additional conditions to the antecedens. Therefore Hempel's own account yields confirmation in both cases. In the first case (we do not know beforehand that a is no raven) this follows from

$$E_1.K_1 = \neg Ra.\neg Ba \models (Ra \rightarrow Ba) = H_{|dom(E)}$$

and in the second case, we have precisely the same implication

$$E_2.K_2 = \neg Ra.\neg Ba \models (Ra \rightarrow Ba) = H_{|dom(E)}.$$

Hence, adding the background knowledge that the test object is no raven does not destroy the (Hempel-)confirmation of H_2. Certainly Hempel spots two points correctly: First, the paradoxical conclusion of the raven example should be embraced, contra (CI). Second, background knowledge plays a crucial role when it comes to explaining the source of the paradox. But while pointing into the right direction, his own theory of confirmation fails to conform to this solution idea of the paradox.

The raven paradox drastically shows how valuable it is to distinguish between evidence and background knowledge and how important it is to formalize this distinction in an adequate way, without running into Hempel's problem. It further shows the problem of monotonicity with regard to evidence and background knowledge: When we happen to know more, confirmation might get lost. Therefore monotonicity in the background knowledge does not see to be a desirable property for accounts of confirmation. The problems of monotonicity and missing evidential relevance will continue to bother us in the next section, when we discuss the hypothetico-deductive approach.

2.3 Hypothetico-deductive confirmation

The hypothetico-deductive approach is one of the oldest and also most intuitive approaches to qualitative confirmation. Roughly, the idea is that from a theoretical hypothesis, we can deduce some predictions with the help of background knowledge. If such predictions are observed, they constitute a confirmation of the hypothesis. Theories whose predictions are observed multiple times and whose predictions never go wrong are better *corroborated* by the evidence than those whose predictions fail to obtain. Thus, hypothetico-deductive confirmation bears a close resemblance to the falsificationist principles of conjecture and refutation. Take, for example, Thomas Young's famous double-slit experiment (1801) that contrasted the hypothesis that light exhibits wavelike behavior and the hypothesis that light is composed of particles. To confirm the wave hypothesis, Young set up the double-slit experiment where a beam of light is shot at a solid and opaque plate that has two open slits in it. Behind the plate, there is a white screen where the light that passes through the slits is recorded. If light is indeed a wave, we expect that wave fronts emerge from each slit, propagate in concentric circles, interfere with each other and yield an interference pattern that is characteristic of a wave. Indeed, when both slits are open, we see such an interference pattern – a pattern of alternating light and dark bands on the screen (see figure 2.1 and 2.2). This observation thus confirmed Young's contention that light exhibits wavelike behavior. The methodology of the experiment is top-down and deductive: *If* light were an electromagnetic wave *then* we would observe interference patterns on the screen. Since we do in fact make those observations, the hypothesis about the wavelike nature of light is confirmed. The wave hypothesis has survived a severe test and an attempt to be falsified – had other results been observed, we would have had to modify the wave hypothesis. On the other hand, the experiment shows that something must be wrong with the classical corpuscular theory.

Formally, the employed scheme of reasoning can be put as

Definition 2.7 *Hypothetico-deductive Confirmation (H-D confirmation): E H-D-confirms H relative to K if and only if (1) $H.K$ is consistent, (2) $H.K$ entails E ($H.K \models E$) and (3) K alone does not entail E.*

In other words, an evidence report confirms a hypothesis if and only if it is a joint implication of hypothesis and background assumptions, if the latter are jointly consistent and if the background knowledge does not entail the evidence. The last condition is adduced in order to avoid that an arbitrary hypothesis is confirmed by the evidence just because the background assumptions already entail the evidence. To see an application of that scheme to Young's double-slit experiment, see 2.1.

In contrast to Hempel's satisfaction criterion, the evidence is now deductively entailed by the hypothesis (and the background knowledge). This model of confirmation exhibits parallels to Popper's falsificationist scheme

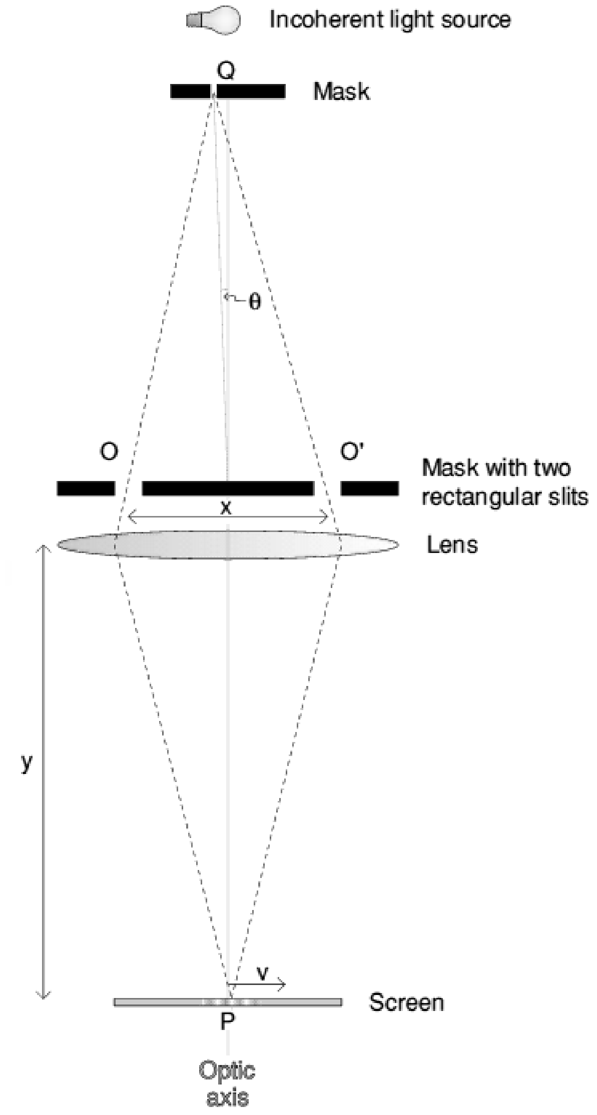

Figure 2.1: The setup of Young's double slit experiment. Source: en.wikipedia.org.

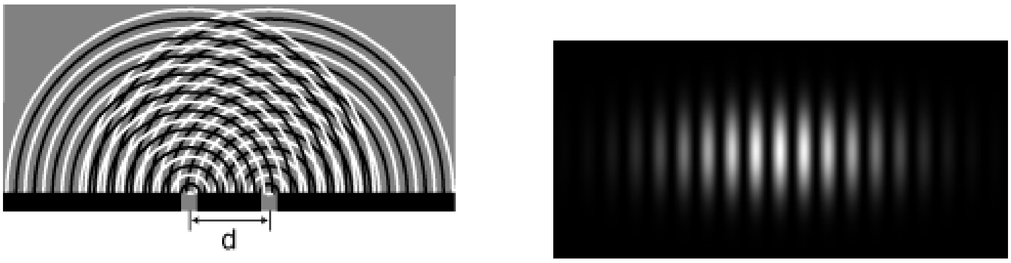

Figure 2.2: The interference pattern of light after passing through the double slit. Source: en.wikipedia.org (left figure), www.paulfriedlander.com (right figure).

Light exhibits wavelike behavior	(Hypothesis)
A beam of light passes through two slits in an opaque plate	(Background assumption)
The light is recorded on a screen behind the plate	(Background assumption)
When sent through two slits, waves exhibit interference patterns	(Background assumption)
An interference pattern is displayed on the screen	(Observation report)

Table 2.1: Young's double-slit experiment, interpreted as a case of H-D confirmation.

of conjecture and refutation:[15] a case of H-D confirmation lends partial, but never full support to the hypothesis, but had another outcome been observed, the hypothesis would have been falsified.

So how does the hypothetico-deductive account of confirmation deal with the raven paradox? Recall that the raven paradox sets up the following problem: If we assume that the raven hypothesis $H = \forall x : Rx \rightarrow Bx$ ('all ravens are black') is confirmed by its instance $E_1 = Ra.Ba$, contraposition yields that $H = \forall x : Rx \rightarrow Bx$ is equally confirmed by $E_2 = \neg Ra.\neg Ba$ (e.g. 'a is a white piece of chalk'). This looks unsound, and many philosophers, as Carl G. Hempel, have tried to argue that it is plausible and desirable that $\neg Ra.\neg Ba$ confirms H. So it might be surprising that in a H-D account, instances of universal conditionals often fail to confirm – in particular, neither $Ra.Ba$ nor $\neg Ra.\neg Ba$ H-D-confirms $H = \forall x : Rx \rightarrow Bx$ relative to empty background knowledge. This raises a number of important questions: Can such hypotheses be H-D confirmed at all? Why is such a behavior not harmful? I think a fair reply can be made. Biting the bullet (=the lack of instance confirmation) does not do any harm since Ba H-D-confirms H relative to the background assumption Ra. In science, several observations are seldom simultaneously made, rather, it seems to be more adequate to describe it as a two-stage observation process: first, property P is checked, then, property Q is checked. As H is intended as a hypothesis about *ravens*, it is sensible to ensure the ravenhood of an object before proceeding to closer investigations. (For scientific properties which are harder to determine than the color of an object, this point is more obvious than in the raven case.) The proposed surrogate confirmation is indeed in line with scientific method: if H is to be tested, we observe some ravens first (Ra) and then examine their color (Ba). So there is nothing to complain about the lack of direct instance confirmation. – Second, I.J. Good (1967, 1968) has argued that instance confirmation does not work in all conceivable circumstances. Assume that we either live in world W_1 with 10 black ravens, 990 crows and no other birdlike objects or we live in world W_2 with 100 black ravens, one non-black raven and 900 crows. Then the observation of a black raven would be more likely in W_2 than in W_1, indicating that we live in W_2 – the world where the raven hypothesis is false. So I conclude that the failure of instance confirmation is not always harmful. So H-D confirmation avoids the problems of an account

[15] See Popper 1963.

that satisfies the Consequence Condition and does not succumb to the raven paradox.

Now I would like to discuss some suggestive objections against hypothetico-deductive confirmation. Note first that H-D confirmation is no complete account of confirmation since existential claims can hardly be H-D-confirmed by observations of individual objects. For instance, Fa does not H-D-confirm the hypothesis $\exists x : Fx$ although this seems to be a clear case of (conclusive) confirmation. But it might be argued that the type of confirmation which H-D confirmation captures – in particular the confirmation of universal claims – is more important than the confirmation of existential claims and that additional criteria should be set up for the latter. Therefore I will neglect this criticism and focus on more direct objections.

Clark Glymour (1980b) believed that H-D confirmation could be led ad absurdum. For this reductio, assume that E and H be contingent and consistent with each other. Relative to the background knowledge $H \rightarrow E$, E can be derived from H. This would meet the crucial condition for H-D confirmation of H by E relative to $H \rightarrow E$. Since E and H were contingent, it is not the case that $H \rightarrow E$ alone entails E. Hence, the H-D conditions are fulfilled, i.e. E H-D-confirms H relative to $H \rightarrow E$. Now let the background knowledge be empty at first and assume that the observed evidence E is true as we normally do. Glymour continues that the (contingent) truth of E implies the (contingent) truth of $H \rightarrow E$ for an arbitrary H since $H \rightarrow E$ is just a *material* conditional. Hence, as the logical consequence of a true statement, $H \rightarrow E$ can be added to our background knowledge. Then we can infer that an arbitrary H is H-D-confirmed relative to true background knowledge ($H \rightarrow E$) by any true evidence E.

I would, however, deny that this is an embarrassing feature of hypothetico-deductive confirmation. To my mind, Glymour misconstrues the relation between background knowledge and evidence: In Glymour's example, the background knowledge is not really *known*, it is *derived* from the evidence. Due to this ad hoc character, it fails to provide a *background* for the evaluation of the evidence. This might look like an untenable relativization of confirmation relations to epistemic states. But all confirmation relations must be evaluated in the light of background assumptions. Glymour commits the fallacy to double-count the evidence: He infers from the evidence E to the assumptions which serve as a background for evaluating whether the same E confirms the hypothesis. Scientific method strictly prohibits this step and maintains that evidence must be counted only once. So it is no wonder that we get strange results. On the other hand, if $H \rightarrow E$ is part of our *a priori* background knowledge, we get completely sound confirmation: A hypothesis is tested by checking its evidential consequences by means of the bridge sentence $H \rightarrow E$. Thus we confidently reject Glymour's conclusion.

A really embarrassing group of objections are, however, the tacking paradoxes. It is possible to tack *irrelevant conjunctions* to the hypothesis H and to preserve the confirmation relation: If H is confirmed by a piece of evidence E (relative to any K), $H.X$ is confirmed by the same E for an arbitrary X that is consistent with H and K. We can easily check the three conditions for H-D confirmation: First, by assumption, $H.K.X$ is consistent. Second, if $H.K \models E$ then also $H.K.X \models E$ because logical implication is monotone with regard to the antecedens. Third, K alone does not entail E because E H-D-confirms H relative to K. Thus, tacking an arbitrary an irrelevant conjunct to a confirmed hypothesis preserves the confirmation relation.[16] It is easy to see that this is highly unsatisfactory: Assume that the wave nature of light is confirmed by Young's double slit experiment. According to the H-D account of confirmation, this also means that the following hypothesis is confirmed: 'Light is an electromagnetic wave and Earth is a disc.' This sounds completely absurd. The problem is pressing and we have to resolve it.

The above problem has a counterpart with regard to the evidence. Tacking *irrelevant disjunctions* to the evidence E equally preserves the confirmation relation: If E confirms a hypothesis H, $E \vee E'$ H-D-confirms the same H for an arbitrary E' unless K logically implies $E \vee E'$. By assumption, $H.K$ is consistent (condition 1) and from $H.K \models E$ it follows that $H.K \models E \vee E'$ (condition 2). And condition 3 of H-D confirmation (K alone does not entail E) was already presupposed. Again, this tacking problem has unacceptable consequences.[17] The hypothesis 'Light is an electromagnetic wave' is H-D-confirmed by the observations in the double-slit experiment (the interference pattern on the screen). Hence, it is also confirmed by the experimental observations *or* the observation that my neighbor's cat is black. This is as absurd as the tacking of arbitrary conjunctions. Both objections exploit the

[16] The problem is mentioned in Glymour 1980b (among other sources). Nowadays, it is present in any contemporary essay about H-D confirmation.

[17] Extensive discussions are given in Gemes 1993, 1998 and more recently in Moretti 2006.

fact that classical H-D confirmation gives no account of *evidential relevance*. All these failures of classical H-D confirmation might lead to the conclusion that the entire approach is hopeless and should be replaced by a refined instance view of confirmation, analogous to Hempel's satisfaction criterion (see Glymour 1980a, 1980b). Nevertheless, some philosophers have undertaken remarkable efforts to rescue the H-D account of confirmation because it so nicely fits our intuitive scheme for confirmation in science. I discuss several attempts to reply to the above challenges and to introduce evidential relevance into a H-D account of confirmation, beginning begin with a proposal by Paul Horwich (1982).

Horwich does not assign a direct role to the background knowledge and believes that a confirmation predicate with only two places can do the job. The idea is thus: An evidence report E confirms a hypothesis if E can be decomposed into conjunctions so that the classical H-D definition is satisfied. One of the conjunctions takes the role of the (H-D-)evidence, the other one the role of the (H-D-)background knowledge.

Definition 2.8 E H-D confirms H according to Horwich *if and only if* E *has a decomposition* $E = E_1.E_2$ *with wffs* E_1 *and* E_2 *so that (1)* E_1, E_2 *and* H *are consistent (2)* $E_1.H \models E_2$ *(3) it is not the case that* $E_1 \models E_2$.

In the above definition, E_1 thus takes the role of the background knowledge in H-D confirmation. Horwich's tricky idea tries to get around explicit consideration of the background knowledge and he might argue for this step by noting that reference to background knowledge does not occur in our everyday usage of the confirmation concept. Moreover, some oddities of classical H-D confirmation disappear, e.g. $Ra.Ba$ now confirms $H = \forall x : Rx \rightarrow Bx$ (due to the decomposition $E_1 = Ra$, $E_2 = Ba$), in contrast to the classical definition. But the tacking paradoxes are not resolved. If $E = E_1.E_2$ confirms H, $E' = E_1.(E_2 \vee E_3)$ will do the job, too, for an arbitrary E_3. Hence, the problem of missing evidential relevance was not solved, but only disguised in different clothes. But the main drawback of Horwich's proposal is a straightforward counterexample. For any contingent E and H, E is logically equivalent to $(H \rightarrow E).E$, $\rightarrow$ being read as a material conditional. Now, we use this observation to confirm H by E with the decomposition $E_1 = H \rightarrow E$ and $E_2 = E$. Obviously, this suffices for classical H-D confirmation – see Glymour's previous objection. So an arbitrary piece of evidence confirms any hypothesis – and Horwich's entire criterion is trivialized.[18] So let us look for other attempts.

In his 1990, John Grimes suggest to replace classical H-D confirmation by 'disjunctive' H-D confirmation. This applies only to truth-functional compounds of atomic well-formed forms, but that is enough for most cases of confirmation. First, any observation report E is transformed into its disjunctive normal form.

Definition 2.9 *A logical formula ϕ is a* literal *if and only if ϕ is an atomic formula or a negation thereof.*

Definition 2.10 *A logical formula ϕ is in* disjunctive normal form (DNF) *if ϕ is a disjunction of one or more conjunctions of one or more literals.*

For instance, Fa, $\neg Ga$ and $\neg Gb$ are literals, but not $Fa.Ga$. They are atomic formulas that might be equipped with a negation symbols. To see the significance of a DNF in informal terms, you might imagine it as the enumeration of all states of the world that make the formula true. The formula $Fa \vee Ga$ has the disjunctive normal form $Fa.Ga \vee \neg Fa.Ga \vee Fa.\neg Ga$, and so on. To cope with the tacking by disjunction paradox, Grimes now suggests that a hypothesis need not entail the entire evidence, but only an element of the disjunctive normal form. If we embed this proposal into our account of H-D confirmation, we obtain

> *Disjunctive H-D confirmation*: Let $E = D_1 \vee \ldots \vee D_n$ be an observation report in its disjunctive normal form. E H-D-confirms H relative to K if (1) $H.K$ is consistent, (2) there is a $k \leq n$ so that $H.K \models D_k$ and (3) K alone does not entail E.

A little example motivates why the disjunctive version of H-D confirmation can indeed capture evidential relevance: Take the hypothesis that all ravens are black ($\forall x : Rx \rightarrow Bx$). We face the 'tacking by disjunction' problem – relative to the background knowledge that a is a raven, this hypothesis is H-D-confirmed by the observations that a is black (Ba) *or* the b is a dove of indefinite color (Db). Now we construe the disjunctive normal form $E_{\text{dnf}} = Ba.Db \vee Ba.\neg Db \vee \neg Ba.Db$. Obviously, the hypothesis makes no claims about whether b is a dove or not so that, and indeed, no disjunctive

[18] See Gemes 1998, 3.

component of E_{dnf} it entailed by $H.K$. Thus the condition of disjunctive H-D confirmation is not satisfied and spurious cases of confirmation due to the tacking of irrelevant disjunctions vanish.

Unfortunately, Grimes' suggestion is not as effective as it seems to be at first sight.[19] Assume that E_1 and E_2 are atomic sentences (e.g. Fa and Fb, a and b being individual constants) which confirm H according to disjunctive H-D confirmation. Assume further that neither piece of evidence implies the other one and that the background knowledge is empty. Then $E_1 \vee \neg E_2$ disjunctively confirms H, and again, non-confirming disjunctions have been tacked to the evidence. To give an example: $H = \forall x : Fx$ is confirmed by $E = Fa \vee \neg Fb$ because the disjunctive normal form of E is $E_{dnf} = Fa.Fb \vee Fa.\neg Fb \vee \neg Fa.\neg Fb$ and the first element of E is entailed by H. But the evidence only tells us that a prediction of H is true or H has been falsified. Definitely, it does not support H. This should not count as confirmation of H so that Grimes' suggestion is seriously inadequate.

After seeing the failure of Grimes' proposal, it might be argued that a principle which seriously restricts the tacking paradoxes would be too strong anyway (Moretti 2006). It is maybe not always the case that tacking irrelevant formulas to the evidence destroys the confirmation relation. However, I am not satisfied with this reply. This exit road is an option in quantitative confirmation theory, e.g. for a Bayesian, but certainly not in qualitative confirmation theory. The point of qualitative confirmation consists in singling out relations of evidential relevance and in adequacy with regard to *actual* cases of confirmation in science. Certainly, those disjunctively tacked pieces of evidence do not fulfil those criteria. And indeed, there are suggestions how to meet those problems.

The source of the tacking problem is the fact that logical implication is insensitive to relations of evidential relevance. When $H \models E$, it does not matter how many sentences we disjunctively add to E, the entailment relation is preserved. Therefore any approach that would like to remedy the problems of H-D confirmation has to develop a theory of relevant entailment, too. We do not want to invoke relevance logic and leave the realm of standard first-order logic because scientific theories are usually formalized in first-order logic. A promising approach in this direction was made by Gerhard Schurz (1991, 1994, 2005). His criterion is based on the *replaceability*

[19] The following objection was raised by Ken Gemes (1993, footnote 4) and appears again, seemingly independently, in Moretti (2004, 19).

of a well-formed formula in the consequens of a logical implication. Such a term is replaceable if we are allowed to replace it by another formula without destroying the logical implication. Consider, for example, the logical implication $Fa \models Fa.Fb$. Obviously, we can replace Fb by Gb, $\neg Hb$ or whatever formula without invalidating the logical implication.[20] To see the difference, note that we cannot replace the formula Fa in the consequens by an arbitrary formula. Now, we would like to mark irrelevant conclusions by the fact that they contain replaceable elements – or as Schurz puts it, predicates that are replaceable on some of their occurrences. A little bit more technically, Schurz distinguished relevant and irrelevant conclusions thus (Schurz 1991):

Definition 2.11 *Assume* $\Gamma \models \phi$. ϕ *is a* relevant conclusion *of* Γ *if and only if no predicate in* ϕ *is replaceable on some of its occurrences by any other predicate of the same arity, salva validitate of* $\Gamma \models \phi$. *Otherwise,* ϕ *is an* irrelevant conclusion *of* Γ.

Definition 2.12 *Assume* $\Gamma \models \phi$. *If* ϕ *is an irrelevant conclusion of* Γ, $\Gamma \models \phi$ *is an* irrelevant entailment.[21]

Schurz's definitions are of special interest to confirmation theory since they naturally apply to the tacking problems. If we tack an arbitrary disjunction to a piece of confirming evidence, the new evidence will merely be an irrelevant conclusion of the hypothesis. Similarly, Schurz comes up with an account that discerns irrelevant premises – the other side of the tacking paradox. I do not spell out the details here and direct the reader to Schurz's publications on that problem (in particular Schurz 1991). Here, it is sufficient to state that the problem can be tackled in a similar way as the problem of irrelevant conclusions. Therefore we get a definition of premise- and conclusion-irrelevant entailment. It is now suggestive to demand that the crucial entailment $H.K \models E$ in the definition of H-D confirmation is neither premise- nor conclusion-irrelevant in order to cope with the objections regarding evidential relevance.

Definition 2.13 E H-D-confirms H *relative to* K according to Schurz *if and only if (1)* $H.K$ *is consistent, (2a)* $H.K$ *entails* E*, (2b)* $H.K \models E$ *is*

[20] To be precise, this is the case as long as the consequens remains consistent.

[21] For the definitions, see Schurz 1991. Schurz gives a syntactic and not a semantic account, focussing on (ir)relevant deductions, but the same definitions can be formulated semantically, as I do it here.

neither a premise- nor a conclusion-irrelevant entailment and (3) K alone does not entail E.

In this definition, the tacking paradoxes vanish. For example, $\forall x : Fx$ is not H-D confirmed by $Fa \vee Ga$ according to Schurz. Similarly, the observation report Fa does not H-D-confirm $\forall x : (Fx.Gx)$ in the modified definition. But on the other hand, $Fa.\forall x : Fx \rightarrow Gx$ is H-D confirmed by Ga. So Schurz's account seems to deal well with the standard objections to hypothetico-deductive confirmation. A problem of that account is, though, the lack of invariance of this account of confirmation under equivalent transformation. (Note that in $\forall x : Fx \models Fa$, the (relevant) conclusion is logically equivalent to $Fa \vee Fa$ which is no relevant conclusion of $\forall x : Fx$.) This violates the equivalence condition for confirmation which is a indispensable element of any account of confirmation: Confirmation should not depend on the way a theory (or an observation report) is formulated and presented. To cope with this problem, Schurz has to make a number of technical modifications, involving an account of *relevant consequence elements.* It would be beyond the scope of this work to discuss those technicalities in detail. Rather, I would like to stress the general point underlying Schurz's work: Elaborating an account of relevant entailment (or relevant deduction) helps to protect H-D confirmation against the menacing tacking paradoxes. A general discussion of the strengths and weaknesses of Schurz's account ensues later, after introducing other approaches.[22]

Ken Gemes's theory of content parts tries to discern irrelevant conclusions along similar lines. Again, the fundamental idea is that relevant logical entailments must not lead to conclusions that contain irrelevant elements. This is captured in the notion of a *content part* (which is something like a relevant conclusion). For instance, $Fa \vee Ga$ would not be a content part of Fa because the element Ga were superfluous. The central concept in Gemes's theory is his account of relevant models (see Gemes 1997, 2006a):

Definition 2.14 *A* relevant model *of a well-formed formula (wff) α is a model of α that assigns values to all and only those atomic wffs that are relevant to α.*

[22] It might be of further interest to note that Schurz's account of relevant entailment is also applicable to other problem, i.e. an account of verisimilitude, deontic logic, etc. Maybe it is due to the versatility of his account of relevant entailment and the fact that a theory of confirmation is only one of several intended application that his account of confirmation has to struggle with some objections which I will make explicit soon.

In other words, there are atomic wffs whose truth value affects the truth value of α. Those wffs count as relevant for α, and a relevant model is a model of α that only cares for those atomic wffs. For instance, the model that assigns 'true' to both Fa and Ga is a relevant model of $Fa \rightarrow Ga$. But the model of $Fa \rightarrow Ga$ that assigns 'true' to Fa, Ga and Ha is not relevant. Based on the definition of relevant models, we get a definition of content entailment – a relation that determines when the consequens in a logical entailment counts as a proper content part of the antecedens:

Definition 2.15 *For two wffs α and β, β is a* content part *of α ($\alpha \models_{cp} \beta$) if and only if (1) α and β are contingent, (2) α logically entails β and (3) every relevant model of β has an extension which is a relevant model of α.*

In other words, β is a content part of α if α logically implies β and if we can extend β to a model of the antecedens α by assigning truth values to further wffs. The content part relation is a means of detecting *irrelevant conclusions*. For instance, $Fa \vee Ga$ is no content part of Fa because the model that assigns 'false' to Fa and 'true' to Ga is a relevant model of $Fa \vee Ga$ but no model of Fa. The content part relation marks such deductions as irrelevant. Similarly, Fa is a content part of $\forall x : Fx$, but $Fa \vee Fb$ is no content part of $\forall x : Fx$ since the relevant model that assigns 'true' to Fa and 'false' to Fb cannot be extended to a relevant model of $\forall x : Fx$. Moreover it is possible to give a *syntactic* version of the content part definition (Gemes 1994). Instead of the third condition in the above definition, we demand that any element of the disjunctive normal form of α is a sub-conjunction of an element of the disjunctive normal form of β. This shows that Grimes was on the right track when paying attention to the disjunctive normal form of the evidence. He just failed to draw the right conclusions.

From the above examples, it is clear that replacing $H.K \models E$ by $H.K \models_{cp} E$ in the definition of H-D confirmation would resolve the tacking by disjunction paradox. But what about tacking by conjunction – the problem of irrelevant premises? The content part definition merely applies to one side of the problem. To this end, Gemes introduces the notion of a natural axiomatization as a set of sentences whose deductive closure is the theory and which are non-redundant content parts of the conjunction of all members:

Definition 2.16 *A set of well-formed formulae T' is a* natural axiomatization *of a theory (a deductively closed set of sentences) T if and only if the* *following three condition are satisfied:*[23]

1. *T is the deductive closure of T'*

2. *every member of T' is a content part of the conjunction of all members of T'*

3. *no content part of any member of T' is entailed by the set of the remaining members of T'.*

For instance, natural axiomatizations of a theory saying that all things are Fs and Gs are both $\{\forall x : Fx; \forall x : Gx\}$ and $\{\forall x : (Fx.Gx)\}$.

This leads to the following refined definition of H-D confirmation, implying a fourth component – the theory to which the hypothesis belongs:

Definition 2.17 E H-D-confirms *axiom A of theory T relative to K* according to Gemes[24] *if and only if*

1. *E is a content part of $A.K$ ($A.K \models_{cp} E$)*

2. *there is no natural axiomatization $N(T)$ of T so that for some set $\mathcal{S} \subset N(T)$, E is a content part of $(K. \bigwedge_{S \in \mathcal{S}} S)$ and A is not a content part of $(K. \bigwedge_{S \in \mathcal{S}} S)$.*

To make this definition more understandable, Gemes claims that

> "only those content parts of T that play a role in the derivation of E can be confirmed by E. In doing so it provides for the type of selective confirmation (or evidential relevance, J.S.) without which H-D (confirmation, J.S.) would [...] be hopeless."[25]

As seen above, taking natural axiomatization into account indeed rules out the classical cases of irrelevant conjunctions. For example, the only natural axiomatizations of $T = \forall x : (Fx.Gx)$ are T itself and the set $N(T) = \{\forall x : Fx; \forall x : Gx\}$. We would like to say that, relative to empty background

[23] I use a slight modification of Gemes's (1993, 483) account.

[24] Gemes (1993, 486) actually suggests a slightly different version in order to meet Glymour's (1980b) criticism, but I have already suggested a rebuttal of this criticism so that we can adopt a less strict formulation.

[25] Gemes 1993, 483-484.

knowledge, T is not confirmed by $E = Fa$ because the component $\forall x : Gx$ is not covered. And indeed, for $\mathcal{S} := \{\forall x : Fx\}$ we obtain that E is a content part of $\mathcal{S}$, but T (which also takes the role of the axiom in definition 2.17) is no content part of $\mathcal{S}$. Therefore, E does not H-D confirm T according to Gemes, in agreement with our desiderata. Similarly, the condition that E be a content part of $A.K$ rules out the cases of irrelevant disjunctions.

Summing up we state that developing notions of relevant entailment and content parts is able to remedy the tacking paradoxes and thus a successful research program to rescue H-D confirmation. Both Gemes and Schurz arrive at important results with the help of their technical tools. How do they compare to each other? Are there deeper problems which are still not recognized? Gemes's theory of content parts has some minor drawbacks which transfer to his account of confirmation, too.[26] For instance, some of Gemes's natural axiomatizations are quite coarse-grained and far from being 'natural'. Let A, B and C denote first-order sentences. Then, the sentence $(A \rightarrow B).(B.C \rightarrow A)$, A, B and C cannot be 'naturally' decomposed into its two conjuncts. These drawbacks are no decisive blow to Gemes's proposal, but certainly, they leave room for improvements.

Schurz, on the other hand, has to give fairly complicated definitions in order to achieve invariance under equivalent transformations of the formulae at stake. Moreover, Gemes (1998, 4-8) has collected some technical objections to Schurz, e.g. $\forall x : Fx$ is not H-D-confirmed according to Schurz by Ga relative to $\forall x : (Fx \rightarrow Gx)$ due to premise irrelevancy. In a similar vein, Schurz must accept that Ba confirms $H = \forall x : Rx \rightarrow Bx$ relative to Ra, but deny that Ba confirms $Ra \rightarrow Ba$ relative to Ra although evidential relevance clearly speaks in favor of confirmation. In particular, the second hypothesis is just a local restriction of the first one.

Hence, although both accounts fare reasonably well in total, they do not give a completely satisfactory solution – if such a solution can ever be attained. H-D confirmation exhibits some parallels to Popperian corroboration – the confirming observation does not prove the hypothesis, but the hypothesis survives a test which aims at the falsification of the hypothesis.[27] For instance, the claim that all ravens are black entails that object a must not be a non-black raven – regardless of whether a is a white piece of chalk or a black raven. Hypothetico-Deductive confirmation is much in line with this

[26] Several minor objections to Gemes are made in Schurz 2005.

[27] See Gemes 1998.

'negative' understanding of confirmation. Indeed, we can see this best in the confirmation of universal sentences (as 'all ravens are black') which can never be proven, but only be supported. In that context, the idea of H-D confirmation is still very widespread in science, and it is important to have an adequate account for it. The next section introduces my own suggestion and tries to solve the open problems by a recourse to falsificationist principles.

2.4 Falsificationist confirmation

Apart from the drawbacks mentioned above, Gemes's and Schurz's proposals are deficient in an important respect. To see the problem in an example, consider the following situation. We would like to test a new antibiotic A against an infection with bacteria of strain S which are, unfortunately, resistant against conventional antibiotics. We set up a clinical trial with a group of infected persons who are given the drug. Besides, we do not set up a control group – the infection might be so dangerous that it would be irresponsible to give a placebo to the other patients. Then we administer the drug to all patients $a_1, \ldots, a_n$ in the treatment group and wait for the results. Indeed, for the first n-1 patients everything works fine and all infection markers soon give negative results. Patient a_n, however, breaks out in a rash before the effect of the drug can be measured. Since a causal connection to the taking of A cannot be ruled out, we stop the treatment. In other words, with regard to a_n, we cannot decide whether the drug is indeed as effective as our hypothesis posits. Still, we have to evaluate the experiment. Do the total observations confirm that antibiotic A eliminates all S-bacteria *and* leads to skin rash?

Clearly, such a claim would stand on very shaky grounds. First, to confirm the effectiveness of A, we should in principle wait for the results of the last patient a_n. Second and worse, the observations do certainly not confirm that taking A always leads to skin rash. To confirm that hypothesis properly, we would have to prolong the experiment and to observe whether the other patients break out in a rash at later time point. But in any case, no responsible medical research report would conclude 'A clinical trial with n test persons has confirmed that antibiotic A is effective against an infection with S-bacteria *and* leads to skin rash'. It is just outrageous to neglect that both claims are not based on the full treatment group, especially since no single patient has exhibited skin rash and lack of S-bacteria. We did not observe a single *instance* of the hypothesis that giving A eliminates S-bacteria and leads to skin rash. Therefore we should (and would) not claim that the composite hypotheses has been confirmed. However, neither the classical nor the refined versions of H-D confirmation agree. Using the plausible formalization[28]

$$\begin{aligned} H_1 &= \forall x\,(Ax \rightarrow \neg Sx) \qquad & K &= Aa_1.Aa_2\dots Aa_n \\ H_2 &= \forall x\,(Ax \rightarrow Rx) \qquad & E &= \neg Sa_1.\neg Sa_2\dots\neg Sa_{n-1}.Ra_n, \end{aligned} \tag{2.1}$$

the combined hypothesis $H_1.H_2$ is H-D-confirmed by evidence E relative to background knowledge K. This result is highly undesirable and does not depend on whether we choose Gemes's, Schurz's or the classical formulation of H-D confirmation. All existing accounts yield the same result (proofs omitted). Nevertheless, the observations $\neg Sa_1, \neg Sa_2, \dots, \neg Sa_{n-1}$ on the one hand and Ra_n on the other hand are completely unrelated so that they should not jointly confirm the composite hypothesis which the single parts clearly fail to confirm.[29] We strongly feel that the evidence should contain at least one instance of $H_1.H_2$, i.e. a patient who, after being given the drug, is free of S-bacteria and develops skin rash. That such an instance is not required makes confirmation of a substantial hypothesis far too easy and opens the door to deliberate manipulation of scientific experiments. To see the latter point in greater generality, imagine that we would like to reason to a foregone conclusion: all objects in a certain group have property P_1 as well as property P_2. Assume now that we find P_1 in some of the objects and P_2 in others. According to all existing variations of H-D confirmation, these observations would confirm that all objects have property P_1 *and* property P_2. This is a fallacy similar to the tacking paradox since in none of our observations, P_1 and P_2 are present in a single object. Therefore, the examined objects should not confirm the *composite hypothesis.* Classical and modern accounts of H-D confirmation open the door to spurious confirmation.

It might be objected that the failure of H-D confirmation should not disturb us too much – there is no completely perfect account of confirmation, and we should simply adjust our intuitions and learn to live with the counterexamples. Such a reply would be fair if the problem were of minor importance and if we could not set up a better account of qualitative confirmation.

[28] Aa denotes that antibiotic A has been given to person a, Sb denotes that S-bacteria are present in the body of patient b, R is the skin rash predicate.

[29] Here, we can spot the problem of irrelevant conjunctions in a new cloak.

But first, we often want to confirm the conjunction of several hypotheses from a single evidence set so that the problem is clearly pressing. Second, we believe that a better account that takes a new, falsificationist approach, is available, and we would like to sketch it now. The new account stands on three pillars. First, a hypothesis is confirmed by checking its predictions. Second, the evidence has to put the hypothesis to a serious test, in other words, had a result different from the actual one obtained, the hypothesis would have been falsified. Those two principles combine hypothetico-deductivist principles with falsificationist philosophy of science – scientific theories can be falsified, but usually not be verified. However, successful predictions that put a theory in jeopardy confirm it. In Popper's own words:

> "Confirming evidence should not count except when it is the result of a genuine test of the theory; and this means that *it can be presented as a serious but unsuccessful attempt to falsify the theory.*"[30]

Third and last, instances of a hypothesis have a distinguished position when it comes to confirmation, as the above example has made clear.

Now we can proceed to the formalization of those principles. First, evidential predictions are logical consequences of a hypothesis H together with background assumptions K. Hence, $H.K \models E$ is a necessary condition for the new account. As we would like to circumvent the problem of tacking by disjunction, we restrict ourselves to relevant entailments and proceed to the stronger condition $H.K \models_{cp} E$. Second, evidence counts as confirming when the conjecture under test has survived an attempt to be falsified, as expressed in the above Popper quote. In other words, if E is to confirm H, $\neg E$ has to falsify H. Usually, this falsificationist idea is formulated in terms of standard logical entailment, $\neg E \models \neg H$, akin to classical hypothetico-deductivism. But we have to be careful – the falsification relation should take into account relations of *evidential relevance*, too. To spot the problem, note that $\neg E \models \neg H$ entails $\neg E \models \neg(H.X)$ for an arbitrary X. Formalizing falsification by means of standard logical entailment does not distinguish between the actual hypothesis under test and any logically stronger hypothesis – both would be falsified even if only the actual hypothesis had a relevant connection to the evidence. To remedy that problem, we reapply the *content part* relation. More precisely, we demand that $\neg H_{|dom(E)}.K$, the restriction

[30] Popper 1963, 37, my italics.

of H to the domain of E plus the background knowledge, be a content part of $\neg E.K$.[31] This amounts to the condition

$$\neg E.K \models_{cp} \neg H_{|dom(E)}.K \tag{2.2}$$

and leads to the following definition of falsificationist confirmation or F-confirmation.

Definition 2.18 ***(Falsificationist Confirmation (FC)):*** E F-confirms H relative to K *if and only if*

- *E is a content part of $H.K$ ($H.K \models_{cp} E$) and*
- *$\neg H_{|dom(E)}.K$ is a content part of $\neg E.K$ ($\neg E.K \models_{cp} \neg H_{|dom(E)}.K$).*

Note in particular that deductively gained instances of a hypothesis usually satisfy (2.2) and F-confirm the hypothesis. And vice versa: if we do not get a full instance of H, we normally fail to F-confirm H.[32] So the falsificationist approach to confirmation assigns a special position to instances of a scientific hypotheses – a point that was already stressed by Hempel (1965) and Glymour (1980a) and that reappeared above.[33]

Now, it has to be seen whether the new account is able to deal both with the classical and the novel objections to H-D confirmation. A main challenge for deductive theories of confirmation consists in the tacking paradoxes that clash with our view that confirmation must not be arbitrarily transmitted (see section 2.3).

The second condition of (FC) ensures that, in the case of irrelevant conjunctions, there are relevant models of $\neg H_{|dom(E)}.K$ that cannot be extended to relevant models of $\neg E.K$. For instance, if $H = \forall x\, Fx$, $X = \forall x\, Gx$, $K = \emptyset$ and $E = Fa$, E should not confirm $H.X$ because it is irrelevant to X. Indeed, $\neg(H.X)_{|\{a\}}.K = \neg Fa \vee \neg Ga$ is no content part of $\neg E.K = \neg Fa$, thus avoiding the undesirable result. Partial confirmation, though, is still possible:

[31] The idea to restrict a hypothesis to the domain of the evidence was introduced by Carl G. Hempel [1945] (1965). Omitting the restriction would not work, because, for a sufficiently general H, (2.2) would never be satisfied.

[32] This is especially pronounced when the evidence is a truth-functional compound of atomic wffs.

[33] It is easy to see that (FC) satisfies the Equivalence Condition which is a basic requirement for all formal accounts of confirmation: if H is logically equivalent to H', then E confirms H relative to K if and only if E confirms H'.

If we add the background knowledge $K = Ga$ (instead of tautologous K), $E = Fa$ F-confirms $H.X$ relative to K. This corresponds to our intuitions that partial confirmation is sound as long as the background knowledge provides the missing piece of evidence that we need for a full instance of the hypothesis. Put another way, partial confirmation counts as F-confirmation whenever the background knowledge covers that part of the predictions of the hypothesis which the evidence does not contain itself.

A corresponding problem for H-D confirmation arises if the evidence is logically weakened, i.e. if irrelevant disjunctions are tacked to the evidence. Assume that a hypothesis H is confirmed by a certain piece of evidence E. If we tack an arbitrary disjunct E' to the evidence, classical H-D confirmation of H remains intact because $\models$ is a transitive relation and $H \models E \models (E \vee E')$. But we do not think that such an $E \vee E'$ is still a relevant prediction of H because E' could be anything. Indeed, the first condition of (FC) requires confirming evidence to be a content part of $H.K$. If an irrelevant disjunction is tacked to the evidence, there will be relevant models of the compound evidence which cannot be extended to relevant models of $H.K$. For instance, if $H = \forall x\, Fx$, $K = \emptyset$, $E = Fa$ and $E' = Gb$, $E \vee E' = Fa \vee Gb$ is no content part of H.[34]

Finally, the behavior of falsificationist confirmation with regard to composite hypotheses avoids the problems of conjunctive confirmation. I would like to come back to the medicine example. Assume that E_1 confirms H_1 and E_2 confirms H_2 relative to K in the H-D sense. In general, conventional accounts of H-D confirmation now affirm that $E_1.E_2$ confirms $H_1.H_2$ relative to K, too.[35] Although this inference looks tempting, we have revealed the problems which come along with this property. Indeed, falsificationist confirmation does not instantiate that scheme in general and thus differs from all previous accounts. In the antibiotics example (2.1), $E = \neg Sa_1. \ldots .Sa_{n-1}.Ra_n$ does not F-confirm $H_1.H_2 = \forall x(Ax \rightarrow \neg Sx).\forall x(Ax \rightarrow Rx)$ relative to $K = Aa_1.Aa_2 \ldots Aa_n$ – the first condition of 2.18 is satisfied, but the second, falsificationist condition is violated. Nonetheless this does *not* rule out the confirmation of conjunctions of independent hypotheses. Consider the

[34] See Gemes 1993.

[35] This is also suggested by Goodman (1983, 71).

following case:

$$\begin{aligned} H_1 &= \forall x\,(Rx \rightarrow Bx) & E_1 &= Ba \\ H_2 &= \forall x\,(Dx \rightarrow Wx) & E_2 &= Wb \\ K &= Ra.Db.(\forall x\,\neg(Rx.Dx)) & & . \end{aligned}$$

The situation is quite analogous to the antibiotics example. Let us interpret H_1 as the claim that all ravens are black and H_2 as the claim that all doves are white. Then the background knowledge K asserts that a is a raven, b is a dove and nothing is both a raven and a dove. This guarantees that H_1 and H_2 are truly independent from each other. Therefore, the observation $E_1.E_2$ that a is black and b is white should confirm the composite hypothesis $H_1.H_2$ that all ravens are black and all doves are white. Falsificationist confirmation agrees and yields confirmation. But if we had omitted the background knowledge that the sets of ravens and doves are disjoint, $H_1.H_2$ would not have been confirmed. And rightly so because a could have been a non-white dove or b could have been a non-black raven. Falsificationist confirmation is thus fine-grained and sensitive to the peculiarities of the background knowledge whereas Gemes's and Schurz's proposals unanimously affirm confirmation in the above case as well as in the medicine example.

By combining deductivist and instantial views of confirmation, (FC) avoids a lot of contentious properties of a purely H-D approach and complies with our intuitions about evidential relevance. Apart from that, (FC) is considerably simpler than other deductive accounts of confirmation. Although the definition of (FC) has some parallels to Gemes's account of H-D confirmation (see, for instance, Gemes 1998), it is clearly more parsimonious: Gemes suggests a criterion which involves the *natural axiomatization* of a theory to which the hypothesis belongs. But first, this introduces a fourth place into the confirmation relation – the theory to which the hypothesis belongs. That is a pretty severe modification. Second, it is open to serious discussion how to fix the notion of a natural axiomatization: As seen in the previous section, Gemes's natural axiomatizations are not very fine-grained and in some cases far from being the 'natural' representations of a hypothesis. On the other hand, Schurz's own suggestion – hypotheses must be represented as conjunctions of their *relevant consequence elements* – also involves complications. Thus, (FC) defines H-D confirmation in a simpler, less contentious and more fruitful definition than the rival proposals: Neither natural axiomatizations nor relevant consequence elements have to be introduced, contributing to the overall attractivity of (FC).

2.5 Summary

Confirmation in science often refers to structural relations between theory and evidence and not always to increase in degree of belief, as probabilistic theories of confirmation suggest. Therefore, in spite of the present popularity of probabilistic approaches, studying qualitative confirmation is indispensable to understand and to reconstruct arguments in a large variety of empirical sciences. We focus on a three-place version of the confirmation relation – evidence confirms a hypothesis relative to background knowledge. But finding a viable account of qualitative confirmation has proven to be a demanding and intricate task. Hempel tried to establish the *satisfaction condition*: the 'development' of a hypothesis for the special experiment is entailed by the observed evidence. But apart from various minor concerns, Hempel's satisfaction criterion runs into great trouble when applied to the paradox of the ravens. This calls our attention to the hypothetico-deductive (H-D) tradition in confirmation theory: evidence confirms a hypothesis when it is deductively entailed by the hypothesis (e.g. when it is a prediction of the hypothesis). The classical version of H-D confirmation surrenders to the tacking paradoxes – tacking irrelevant conjuncts and disjuncts to hypothesis and evidence does not destroy the confirmation relation. The more refined proposals of Gemes and Schurz are able to resolve those problems. However, both accounts give an unintuitive and coarse-grained treatment of confirmation of composite hypothesis. I have argued that my own proposal – falsificationist confirmation – gives a convincing answer. Since that criterion circumvents the tacking paradoxes, too, and can be formulated in a way that is clearly simpler than the definitions of Gemes and Schurz, it is the most hopeful candidate in a series of attempts to rescue the hypothetico-deductive account of confirmation in science.

We may now wonder whether those basic accounts of confirmation can be extended to an account of the confirmation of entire theories. Furthermore, we are interested in discussing the problem of holism and resuming the discussion about the role of the background assumptions in confirmation theory. These problems are the main topics of the next chapter.

Chapter 3

Theory Confirmation

3.1 The holistic challenge

The idea that isolated scientific hypothesis can ever be confirmed was threatened by the holistic challenge, forcefully articulated by Duhem (1914) and later put forward by Quine (1961). The basic idea is very simple: When falsifying a hypothesis by means of an observation, we rely on more factors than the observation alone. We require auxiliary assumptions to derive actual predictions from a (theoretical) hypothesis. Only with the help of such auxiliary assumptions we are able to make testable predictions. Duhem claimed that modern experiments in physics do not only have effect on the special hypothesis under scrutiny: Thermodynamics, mechanics and electrodynamics may interact in a single experiment. In testing a sufficiently complex hypothesis many auxiliary claims from other areas of physics are employed. If such an experiment has a negative outcome, the question arises what has actually been falsified – the hypothesis under test or one of the auxiliary claims?

> "if the predicted phenomenon is not produced, not only is the questioned proposition put into doubt, but also the whole theoretical scaffolding used by the scientist; the only thing experience teaches us is that, among all the propositions which helped to predict the phenomenon and to verify that it has not been produced, there is at least one error; but where the error lies is just what the experiment does not tell us."[1]

[1] The author's translation of Duhem 1914, 181. The original passage reads: "si le phénomène prévu ne se produit pas, ce n'est pas la proposition litigieuse seule qui est mise en défaut, c'est tout l'échafaudage théorique dont le physicien a fait usage; la seule chose

Hence, a negative outcome of an experiment does not force us to say that the hypothesis under test has been falsified; instead, we could also claim that one of the auxiliary assumptions is false. More generally, the negative outcome of the experiment only tells us that *one* of the assumptions – either the main hypothesis or a background assumption – must be dismissed. It remains silent on which of these assumptions is to be blamed. More generally, Duhem argues that it is impossible to submit single hypotheses to an empirical test; auxiliary assumptions and theoretical background always comes into play.[2] Duhem draws the following, famous conclusion:

> "To seek to separate each of the hypotheses of theoretical physics from the other assumptions on which this science rests, in order to subject it in isolation to the control of observation, is to pursue a chimera."[3]

That no single hypothesis can be falsified and that predictive failures cannot be ascribed to single hypotheses constitutes the thesis of falsificational holism:

> *Falsificational Holism (FH)*: Observations only falsify entire theories, not individual parts thereof.

This thesis makes a *logical* point about the mechanics of falsification, but it can be extended to an *epistemological* point, too: Then we do not only claim that can no hypothesis can be falsified without invoking auxiliary hypotheses, but also that hypotheses cannot be *confirmed* in isolation. Instead, only groups of hypotheses or entire theories are confirmed by observations. This is the tenet of confirmational holism.

que nous apprenne l'expérience, c'est que, parmi toutes les propositions qui ont servi à prévoir ce phénomène et à constater qu'il ne se produisait pas, il y a au moins une erreur; mais où gît cette erreur, c'est ce qu'elle ne nous dit pas."

[2] This argument can be further reinforced by Kuhn's (1962) famous point that all observation is theory-laden – there is no theory-independent observational language. Hence, the background theory is not only present in the auxiliary assumptions, but also in the observation itself. Therefore, no single hypothesis faces the test of experience alone, independent of the theory. I will return to that point later.

[3] The author's translation of Duhem 1914, 303. The original passage reads: "Chercher à séparer chacune des hypothèses de la Physique théorique des autres suppositions sur lesquelles repose cette science, afin de la soumettre isolément au contrôle de l'observation, c'est poursuivre une chimère."

> *Confirmational Holism (CH)*: Observations only confirm entire theories, not individual parts thereof.

Confirmational holism was endorsed by Willard Van Orman Quine, on the grounds that Duhem had previously argued for falsificational holism. From the fact that no contrary experience can falsify a single hypothesis Quine concludes that there is a wide variety of choices which beliefs to maintain and which to abandon in the face of experience. Quine argues that no relations of evidential relevance decide over confirmation and disconfirmation – instead we make the kind of adjustments that keep our entire system of beliefs in balance:

> "No particular experiences are linked with any particular statements in the interior of the field, except indirectly through considerations of equilibrum affecting the field as a whole. [...] Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system."[4]

Hence, Quine concludes that no stand-alone empirical statement can be refuted or supported by experience – our claims about the external world do not face experience in isolation, but merely as a collective.

The distinction between (FH) and (CH) is not always explicitly made, and in the holism debate these points are often conflated. In his insightful discussion, Morrison (2008) illuminates the havoc that is wreaked by not disentangling those tenets. Quite obviously, confirmational holism is logically stronger than falsificational holism since falsification and verification are only particular ways of (dis)confirming a hypothesis. Confirmational holism is a major worry for confirmation theory: If that claim were true, we would not be able to say whether (and to which degree) a piece of evidence bears the relation of evidential relevance to a single hypothesis. Furthermore, we would have problems to model the confirmation of scientific claims in the history of science and to acknowledge the value of crucial experiments, etc. Endorsing a confirmational holism would concede that we are not able to understand why scientists argue and experiment as they do and why they are so successful at doing so. But the holistic challenge is not only restricted to philosophy of science – Quine (1961) uses the holistic argument against the reductionist enterprise of the logical positivists and the analytic/synthetic distinction. To

[4] Quine 1961, 43.

see this more clearly, the auxiliary assumptions which connect the theoretical to the observational vocabulary of a scientific language are sometimes called ‘meaning postulates’ and awarded an ‘analytical’ character, bridging the gap between theory and observation. Quine objects, however, that a negative test of a theoretical hypothesis could not only lead to rejection of that hypothesis, but equally to a rejection of the ‘analytic’ meaning postulates, putting in jeopardy the analytic/synthetic distinction. Thus the problem is not only restricted to philosophy of science – it has implications for philosophy of language and in modern applications even for philosophy of mathematics (Colyvan 2001). I do not want to discuss these issues in a deeper way, but nevertheless, enumerating these implications helps to make clear that the holistic challenge is enormously relevant and that a successful confirmation-theoretic reply to confirmational holism is highly desirable.

A natural question concerns the relationship between the weaker falsificational holism and the stronger confirmational holism. Obviously, the former is implied by the latter since falsification and verification are special forms of (dis)confirmation. But which of the two claims is actually supported? Duhem has argued for falsificational holism: Modern physical experiments involve auxiliary hypotheses from various physical theories, thus making it impossible to allocate the error to a specific hypothesis or theory. It is impossible to derive predictions from a theoretical hypothesis and to falsify it without taking theoretical background or auxiliary theories into the boat. Duhem’s observation is certainly correct, hence falsificational holism seems to be well supported. But what about confirmational holism? If we look into binoculars and see a black raven, this observation seems to be confirmationally relevant to the hypothesis that all ravens are black. It is much harder to see how it is relevant to the hypothesis that the binoculars are working properly although this hypothesis is required for claiming that a black raven has really been observed. This is, for instance, captured in Hempel’s satisfaction criterion – the observation of a black raven is an *instance* of the hypothesis that all ravens are black, but not of the hypothesis that the binoculars work fine. Hence, the relationship between (FH) and (CH) depends on the specific way an account of confirmation spells out the relation of evidential. Only in a very basic, primitive formulation of hypothetico-deductivism,

> (2HD): E H-D-confirms T if and only if T logically implies E ($T \models E$).

confirmational holism follows from falsificational holism (see Morrison 2008). Assume that T is an empirical theory (or a complex of theories) consisting of various hypotheses. When T implies E but no proper part of T does, only the theory T as a whole is confirmed according to (2HD). Confirming evidence is entailed by the predictions of the theory, and confirmation is equated with successful survival of an attempt to falsify the hypothesis. Using such a definition of confirmation, falsificational holism entails confirmational holism. However, we have seen in the previous chapter that such a primitive version of H-D confirmation gives a poor account of evidential relevance and succumbs to the tacking paradoxes. Therefore we are well advised to reject (2HD).

Furthermore it has been argued (e.g. Sober 1999) that confirmational holism is a descriptively inadequate and scarcely understandable claim. In empirical experimentation, there is a crucial distinction between hypotheses *in use* (the auxiliary assumptions) and hypothesis *under test* (the target of inquiry). A lot of experimental practice is concerned with *isolating* single hypothesis and putting them to test, whereas other hypothesis only play an auxiliary role. It is hard to imagine experimental practice that dismisses that distinction and such a step would not correctly describe scientific activity. There must be something about confirmation that allows scientists to ascribe confirmation to specific claims while keeping others fixed. If we endorsed confirmational holism, this distinction would completely get lost and a lot of experimental practice could not be properly understood. Therefore we should try to get around (CH).

The natural move that accommodates those worries consists in modifying the confirmation predicate from a two-place predicate (hypothesis and evidence) to a three-place predicate (hypothesis, evidence and background knowledge). This allows a natural distinction between hypotheses under test and hypotheses in use, the latter (=the auxiliary assumptions) being a part of the background knowledge rather than of the tested hypothesis. Thus, they are clearly separated from the hypothesis itself and single hypotheses can be confirmed relative to a body of background assumptions. It can further be asked whether accounts of confirmation always allow to interchange hypothesis and background assumptions while preserving the confirmation relation. That would be a major obstacle for a satisfactory reply to (CH) since such an account would not distinguish between hypotheses in use and hypotheses under test. Put the other way round, the less such an interchange is possible, the more clearly do those accounts outline a relation of evidential relevance.

Indeed, with respect to H-D confirmation, Gemes's and my own falsificationist proposal do not allow an arbitrary exchange of auxiliary hypotheses and hypotheses under test. For a full reply to the holistic challenge, it is, however, desirable to quantify the degree of support which the evidence lends to the various hypotheses in use and under test. In particular, it would be desirable to show, however, that auxiliary hypotheses gain less support from a confirming piece of evidence than the main hypothesis under test. This asks for a quantitative framework (see Strevens 2001, 2005, Fitelson and Waterman 2005a, 2005b) and transcends the power of qualitative confirmation theory. However, some work can also be done in a qualitative framework. To be sure, the confirmation of single hypotheses requires auxiliary assumptions, but we might try to show that theories can be confirmed as a whole, *without recourse to theory-independent auxiliary hypotheses.* We have already seen how parts of a theory can help to confirm other parts of the theory and vice versa. We might now examine the relation between confirmation of single elements of a theory to the confirmation of the entire theory. This project was pursued by Clark Glymour in his development of 'bootstrap confirmation' and it constitutes the subject of the rest of this chapter.[5]

3.2 Theory and evidence

The account of bootstrapping devised by Clark Glymour can be motivated from two different sides. The one side is closely related to the holistic challenge outlined above, the other side (which is often neglected) stems from Kuhn's argument about the theory-ladenness of observation. Let us begin with the first, better known motivation. One of the fundamental problems of epistemology in general and confirmation theory in particular consists in the question how observational data are able to affect the epistemic status of high-level theories which are framed in a more theoretical vocabulary. We have seen that auxiliary hypotheses connecting those two layers of scientific description usually take that task. More precisely, in the confirmation of a scientific theory or parts thereof, we often see intricate moves involving large parts of the theory. This seems to involve a vicious circle – to confirm a theory has to build on parts of the theory itself. In his 1980a, Clark Glymour, however, argues that such arguments for the confirmation of a theory are

[5] See also Glymour 1975 for an analysis of the relationship between evidential relevance and the holistic challenge.

widespread in science and that they are not viciously circular: An argument that employs parts of a theory in order to confirm the very same theory is sound as long as the evidence on which the argument draws puts the theory at risk.

> "Large parts of a theory may be involved in confirming, from given evidence, any of its hypothesis. But it is not true that all of a theory's hypotheses are equally confirmed or disconfirmed (with respect to the theory) by given evidence. [...] We may very well trace conflicts to some special set of claims of a theory makes and dispense with them."[6]

To this end, Glymour devises a formal criterion of theory confirmation, the *bootstrap criterion*. When a conflict between the observations and the theory occurs, the bootstrap criterion is supposed to show which parts of a theory are to blame for the failure and which remain intact. There is a kernel of truth in confirmational holism – we often want to confirm theories as a whole and not just parts thereof. But this does not entail that all of the elements of a theory are confirmed to an equal degree by an observation that confirms the theory according to an 'intuitive' judgment. Glymour counters, too, that embracing a holistic position makes it impossible to understand sophisticated scientific arguments and the ability of scientists to modify exactly those parts of a theory that lead to problems.[7] In particular, Glymour believes that the confirmation of single hypotheses can add up to confirmation of an entire theory where no auxiliary hypotheses are required – more on this later.

Now I would like to introduce the second motivation which is, however, not given by Glymour himself. A major discovery in the history of science that had a deep impact on philosophy of science, too, was the theory-ladenness of observation and evidential statements (Kuhn 1962). The logical positivists pursued the project of clearly separating an observational and a theoretical vocabulary. Sentences in the observational and the theoretical vocabulary were connected by a set of 'bridge principles', 'meaning postulates' or 'coordinating principles' which were often supposed to have analytic status. Thomas Samuel Kuhn (1962), however, found a variety of examples

[6] Glymour 1980a, 151-52.

[7] See Glymour 1980a, 45.

from history of science where such a neat separation did not exist. Scientists stated the observed evidence in the vocabulary which the prevalent theory – the theory they endorsed themselves, the *paradigm* in which they thought – imposed on them. Take for example, the case of a pendulum. In the Aristotelian tradition, the oscillation of a pendulum was seen as a constrained fall: the body eventually moved from a higher to a lower position. In the Renaissance, Galilei conducted quite the same experiments, but saw different things: the energy which the swinging stone gained by moving downwards was transformed into the impetus that displaced the stone to the amplitude again. So where the Aristotelians saw a directed process, Galilei saw a symmetric oscillation that could in principle go on for indefinite time.[8] These observations led Galilei – but not the Aristotelians – to an argument for the independence of mass and rate of fall. The difference between their observations can be traced back to two different paradigms of motions, or so Kuhn argues. Similar examples can be found, in various scientific disciplines (e.g. Lavoisier observed oxygen where Priestley had seen dephlogisticated air). Hence, there are no neutral observations that are independent of the paradigm in which they are made.

> "None of these remarks is intended to indicate that scientists do not characteristically interpret observations and data. [...] But each of these interpretations presupposed a paradigm. [...] In each of them the scientist, by virtue of an accepted paradigm, knew what a datum was, what instruments might be used to retrieve it, and what concepts were relevant to its interpretation."[9]

Here Kuhn stresses the indispensability of a theoretical framework, a paradigm, for interpreting data. Whereas a lot of philosophers in the positivist tradition have affirmed that theories are simply human interpretations of given data,[10] Kuhn answers this question in the negative.

> "The operations and measurements that a scientist undertakes in the laboratory are not 'the given' of experience. [...] The measurement to be performed on a pendulum are not the ones relevant to a case of constrained fall."[11]

[8] See Kuhn 1962, 118-120.

[9] Kuhn 1962, 122.

[10] See Kuhn 1962, 126.

[11] Kuhn 1962, 126.

Hence, scientists holding different theories and paradigms would also state the evidence in a different way. In other words, evidence takes a different meaning in different theories and it cannot play the role of a neutral arbiter between different theories: An observation report is itself entrenched in one theory and alien to the other. Observations are not neutral, but theory-laden. So in assessing a hypothesis and whether it is confirmed by the evidence, it is important to take the theory into account in which the hypothesis and the evidence are situated. This equally transfers to groups of hypotheses and sub-theories. Of course, we are also interested in finding out which of two competing paradigms or general theories is more successful in a specific discipline. Since observations cannot directly decide the matter and since there is no external, theory-independent background against which they could be measured, we need a formal framework that models *how evidence fits into a general theory*, without recourse to external background assumptions.[12] This is, as we will soon see, close to the principal idea of bootstrapping which models deductive moves from the evidence plus parts of a theory to other parts of a theory.

Of course, the theory-ladenness of observation is still a contentious issue in philosophy of science. Thomas Kuhn was a historian of science and it is questionable whether the lack of a theory/observation separation in many episodes from the *history* of science should convince *philosophers* of the general impossibility of such a separation: Maybe the scientists of earlier centuries were just careless and lacked education in philosophy of science so that their failure to separate theory and evidence should not lead us to the conclusion that such a separation is impossible. But Kuhn certainly describes an important problem for science and his argument has been enormously influential in philosophy of science. In the search of descriptive accuracy it is mandatory to accommodate Kuhn's observations in an account of theory confirmation as far as possible. Indeed, Glymour assumes that the auxiliary assumptions are part of the same theory where the hypothesis is taken from:

> "[...] the bearing of evidence is sensitive to changes of theory [...]. For in considering the relevance of evidence to hypothesis, one is ordinarily concerned either with how the evidence bears on a hypothesis with respect to some *accepted* theory or theories, or else one is concerned with the bearing of the evidence on a hypothesis

[12] At one point, Glymour seems to endorse a similar point, see Glymour 1980a, 121.

> with respect to a definite theory containing that hypothesis. In the latter case, the issue is how well the theory is confirmed *with respect to itself.*"[13]

Again, the latter case is typical of a situation where all evaluations of a theory or paradigm can only refer to parts of the theory itself. Thus we encounter cases where piecemeal confirmation of a theory proceeds merely by means of theory-internal auxiliaries whereas we do not have theory-external background knowledge. The various accounts of instance and H-D confirmation do not answer this challenge since they are only concerned with the confirmation of single hypotheses relative to a specific set of background assumptions which is clearly separated from the hypothesis under test. But in looking for an account of theory confirmation, we would like to test a hypothesis relative to the theory to which it belongs, and in general, we do not have a means of separating the content of the hypothesis under test from the theory in which it is embedded.

Showing how claims of a theory are tested against the background of a joint, coherent theory is the core of Glymour's project which is named *bootstrap confirmation* – claims of a theory are tested against the theory itself. Therefore the name 'bootstrap confirmation' – by taking the theory itself as the background of the confirmation relation, one tries to 'pull oneself up by one's own bootstraps.' Consequently, an (axiomatizable) theory is confirmed as a whole if and only if any of its axioms survives an evidential test against the other axioms of the theory. A pretty and concise reconstruction of Glymour's main idea that dismisses the technicalities and builds on an elementary (dis)confirmation predicate is given in Douven und Meijs 2006:

Definition 3.1 *Let $T = Cn(A_1, \ldots, A_n)$ be a finitely axiomatizable theory. E* bootstrap-confirms *T if and only if T and E are consistent and for all $i \in \{1, \ldots, n\}$:*

1. *There is a $T' \subset T$ so that (a) E confirms A_i relative to T' und (b) there are possible (but not actual) observations E' so that E' disconfirms A_i relative to T'.*

2. *There is no $T' \subset T$ so that E disconfirms A_i relative to T'.*

[13] Glymour 1980a, 121. Emphasis in the original.

This definition illuminates an important feature of bootstrapping: the theory itself or a subset thereof can be used in confirming one of its axioms as long as the axiom under scrutiny is put at risk. If this is possible for all axioms of the theory, we achieve bootstrap confirmation of the theory as a whole and thereby unrelativized confirmation of the entire theory. (This is again a concession to Kuhn's point about the theory-ladenness of observation – all possible auxiliary assumptions are themselves embedded in a theory.)

It remains open how the three-place confirmation predicate used in the above definition is to be explicated. Glymour doubts that H-D confirmation can do the job because sophistication is required to make H-D confirmation immune to the tacking paradoxes and related objections (see Glymour 1980b).[14] Moreover, Glymour believes that hypotheses are confirmed by producing instances of them, and contrary to H-D confirmation, these instances are gained from the evidence with the help of elements of the theory. The contrast can be sketched thus:

> "The H-D account looks chiefly at [...] deductive moves from theory to evidence [..]. The new [bootstrap, J.S.] account looks chiefly [...] at deductive moves from evidence plus theory to other theory."[15]

Therefore Glymour phrases his bootstrap confirmation with the help of a predicate that models how instances of a hypothesis are derived from the evidence – in the same way as Hempel.[16] Indeed, Glymour adopts a minor modification of Hempel's satisfaction criterion as his elementary criterion of confirmation. Here, the theory itself is used in a non-redundant way in order to derive instances of a hypothesis that is a part of the theory. We can sum up Glymour's basic idea thus:

> "Hypotheses are tested and confirmed by producing instances of them; to produce instances of theoretical hypotheses one must use other theoretical relations to determine values for theoretical quantities; these other relations are tested in turn in the same way. Ideally, we might hope for bodies of evidence that permit each hypothesis to be tested independently."[17]

[14] Contrary to Glymour, I believe that H-D confirmation is a perfectly proper elementary confirmation predicate. We will come back to that point later.

[15] Glymour 1980a, 168.

[16] See Hempel [1945] 1965 and the previous chapter of this book.

[17] Glmyour 1980a, 52.

Some remarks are due.

1. Glymour admits that many textbook examples of scientific reasoning correspond more to the hypothetico-deductive scheme than to bootstrap confirmation although H-D confirmation has well-known problems with evidential relevance. Glymour gives the following reasons: First, textbook writers tend to simplify methodic subtleties in order to make the confirmation of a theory more understandable. Complex bootstrap arguments are simplified into hypothetico-deductive derivations.[18] For example, Newton's laws of motion and his theory of gravitation do not make predictions about the movement of the planets in the strict sense. For having that, we would have to know the total force acting on the planet. Hence, Kepler's laws cannot confirm Newton's law of gravitation in the hypothetico-deductive sense. But Kepler's laws and the three Newtonian laws jointly entail that a force directed to the sun acts on every planet and that this force is proportional to the inverse square of the planet radius ($F \sim \frac{1}{r^2}$). This dependency is thus a special instance of the general law of gravitation which is derived from Kepler's laws with the help of Newton's laws of motions.[19] This is then a classical case of bootstrap confirmation.

Furthermore, there are often clear expectations in the scientific community what a theory should account for. Observation reports that fall into the 'intended domain' of a theory are automatically relevant for it, and vice versa. For example, elementary theories of matter (as Bohr's early quantum theory) should also determine the spectra of various chemical elements as hydrogen. So we do not have to wonder why Bohr derived spectral series in the hypothetico-deductive way: it was clear beforehand that those observations would be relevant to the theory. When such expectations exist, the classical H-D account gets rid of his most salient problem: the lack of an account of evidential relevance. Therefore Glymour concludes that bootstrap confirmation occurs whenever there are no clearly established, intuitive criteria of evidential relevance. In such a situation the main vice of H-D confirmation – the lack of evidential relevance criteria as exemplified in the tacking paradoxes – would come out clearly. Hence, bootstrap arguments are especially important when novel theories are introduced or when theories are extended to novel fields of application. They are less popular in normal, puzzle-solving science in established fields of research.

[18] See Glymour 1980a, 170-71.

[19] See Glymour 1980a, 169.

2. Glymour's bootstrap confirmation does not give a complete theory of confirmation – it must be supplemented by an elementary criterion of confirmation (confirmation of a single hypothesis with regard to a certain background). Glymour opts for a modification of Hempel's satisfaction criterion, but in principle, he does not rule out the use of another criterion.[20] On the one hand, he stresses that the other axioms of a theory are used in order to derive an instance of the hypothesis – this speaks in favor of a hypothetico-deductive approach. On the other hand, the idea that instances derived from the evidence confirm a hypothesis is well-entrenched in Glymour's position and he is moreover sceptical of the H-D account's ability to account for relations of evidential relevance. But again, in principle, any feasible account of elementary confirmation could be used, so a suitably refined H-D criterion might do the job, too. But not only qualitative, also quantitative theories could play a role here – see Douven and Meijs (2006) for a Bayesian account of bootstrap confirmation.

3. In Glmyour's original account, the confirmation relation between evidence and theory is invariant under the choice of an axiomatization. Theories are just assumed to be deductively closed sets of sentences. On the one hand, this is very attractive since we need not care for a particular axiomatization. But first, this move leads into technical problems as we will see later. Second, scientists rather think about a theory in terms of a coherent network of natural regularities than in terms of a deductively closed set of sentences.[21] This suggests that there are natural axiomatizations of a theory – some that capture the intuitive regularities and others which do not. Therefore bootstrapping might be relativized to a particular axiomatizations of a theory. This approach is pursued in my own and Ken Gemes's (2006b) recent accounts to bootstrapping.

4. Glymour's account of confirmation may appear circular because the background assumptions against which an axiom of the theory is tested consists of a (possibly improper) subset of the theory. Critics may use this to argue that no axiom of the theory is tested truly independently, but only relative to the rest of the theory. Therefore, bootstrap confirmation may appear to be circular. – To my mind, this objection neglects three points: First, the charge of circularity can be rejected because the evidence has to put the theory *at risk* – this was the content of subclause 1b of definition 3.1. Second,

[20] See Glymour 1980a, 127.

[21] See Christensen 1983, 479-480.

there is rarely theory-independent and neutrally describable evidence that is able to decide between different theories and research programs. Some physical theories, e.g. cosmological theories are so fundamental and theory-laden that it would be impossible to confirm such theories relative to a theory-neutral background. That was the point of Kuhn's famous dictum about the theory-ladenness of observation. Instead, researchers attempt to insert pieces of evidence into a coherent picture. But some piece of evidence might coherently fit into a certain theoretical picture and fail to fit into another one, and this is captured by bootstrap confirmation. Third, the *coherence* of a theory contributes a lot to its acceptance and its epistemic status. We prefer a coherent system of beliefs to a incoherent system of beliefs. Clearly, bootstrap-confirmation is coherence-conducive: If a sentence that lacked coherence with the rest of the theory were tacked to a theoretical hypothesis, we would not be able to confirm it relative to the theory because the theory would not provide a bridge between the evidence and that particular hypothesis. The statement would thus stand by itself and disparate from the rest of the theory. Hence, incoherent systems are hard to bootstrap-confirm. In particular, since the elementary confirmation predicate is supposed to take care for evidential relevance, 'irrelevant' or incoherent axioms would quickly be detected. To see the closeness to coherence in greater detail, note that we are usually not able to derive instances of all elements of a theory from the evidence because the relationships between the hypotheses and the evidence are not tight enough. Therefore we have to decompose the theory into several axioms and to 'bootstrap up' the evidence with the help of the rest the theory. Thus only internally coherent theories can be bootstrap-confirmed. We may thus obtain a more modest understanding of what bootstrap confirmation amounts to.

5. Glymour has an original idea how to connect the falsifiability of a theory to bootstrap confirmation and how to point out that falsifiability is an epistemic virtue. We all presuppose that good theories are testable in more than one way, i.e. that there is more than one method to check its predictions for agreement with the observed data. In the Popperian tradition, this is a requirement of scientific method in order to ensure that no pseudo-scientific or ad hoc theory is maintained for a long time. Conversely, when a theory survives a multiplicity of independent tests, it is better confirmed than if only few tests speak for it. Analogously, there can be several ways to bootstrap-confirm a hypothesis: For instance, we compute the value of

certain quantities with the help of theoretical claims in order to test an axiom of the theory that predicts a certain value. The more different ways exist to calculate this value, the less likely it is that mistakes in the various axioms of the theory cancel out each other and erroneously detect confirmation. In other words, multiple falsifiability of a theory protects against erroneous confirmation.[22] Hence, varied evidence and multiple testability come out as virtuous properties in the bootstrap account of confirmation. Glymour gives an example from the history of science, too: Kepler's First Law – planet travel around the sun in ellipses – was not established as a result of accommodating the observed data to a certain hypothesis. Rather, it emerged as a result of the data together with another law – Kepler's Second Law that was confirmed by other data.[23] In other words, for confirming Kepler's First Law the Second Law had to be presupposed. But obviously, the Second Law presupposes the First Law. So sceptics might have raised the suspicion that both laws were wrong but connected in a way that the errors canceled out and the concordance with the data was ensured. It was only after Galilei's discovery of the four Jupiter satellites that it became possible to confirm Kepler's Second Law without explicitly asserting that planets travel around the sun in ellipses. This opened a further way to check the concordance of theoretical claims and observed data and an opportunity to rule out that the errors just canceled out. Thereby the sceptical doubts were eliminated. Glymour finishes:

> "[...] it seems unlikely to me that the development and testing of any complex modern theory in physics or in chemistry can be understood without some appreciation of the way a variety of evidence serves to separate hypotheses."[24]

3.3 Bootstrapping under fire

Glymour's original formulation of bootstrapping opts for Hempel's satisfaction criterion as an elementary criterion of confirmation. In two influential articles, David Christensen (1983, 1990) has pointed out that (a) bootstrap confirmation is too gullible (i.e. that it cannot be a sufficient criterion for

[22] See Glymour 1980a, 139-40.

[23] Kepler's Second Law asserts that the area that a planet sweeps out in a fixed time interval is equal for each set of points of the planet's orbit.

[24] Glymour 1980a, 141.

$A_1 = \forall x : Lx \equiv Zx$	'All Zoroastrians and only them have eternal life.'
$A_1^* = \forall x : \neg Lx \equiv Zx$	'All humans except Zoroastrians have eternal life.'
$A_2 = \forall x : Sx \rightarrow Zx$	'All sudra-wearing humans are Zoroastrians.'
$A_3 = \forall x : Px \equiv Zx$	'All Zoroastrians and only them pray to Ahura-Mazda.'

Table 3.1: The axioms of T and T^*.

theory confirmation) and (b) that it is unable to discriminate plausible and unacceptable cases of confirmation, just because they have the same syntactical structure. From that, Christensen draws the general conclusion that no account of confirmation that considers theories to be just deductively closed sets of sentences runs into the evidential relevance problem. Let us consider his objections.

The thesis that bootstrap confirmation is too gullible can be supported by means of some examples. Take, for instance, the theories $T = \{A_1, A_2, A_3\}$ and $T^* = \{A_1^*, A_2, A_3\}$. Obviously, the two theories T and T^* are inconsistent with each other. But the difference is not 'empirically significant' – it entirely rests in the metaphysical question whether Zoroastrians or Non-Zoroastrians have eternal life. It would therefore seem strange if one of the metaphysical claims A_1 or A_1^* were confirmed by normal, empirically significant evidence. But this problem occurs with regard to bootstrap-confirmation. Take an evidence of the form $E = Pa.Sa$ ('person a wears a sudra and prays to Ahura-Mazda'). In both theories, this innocent observation seems to entitle us to infer to the religion of a, but certainly not to whether a will have eternal life or not. Unfortunately, on Glymour's original account that uses Hempel's satisfaction criterion, E bootstrap-confirms both A_1 and A_1^* with respect to T (T^*). First, E allows us to derive with the help of A_2 that a is Zoroastrian. This is uncontroversial. Second, the sentence $\forall x : Px \equiv Lx$ is part of the deductive closure of T in the same way that the sentence $H = \forall x : Px \equiv \neg Lx$ is part of the deductive closure of T^*. Hence we may use it in the bootstrap confirmation of A_1 respectively A_1^* and we indeed see that the actual evidence entails $La.Za$ respectively $La.\neg Za$. This logically implies the development of A_1/A_1^* for the object a. E is then a confirming *instance* of A_1/A_1^* that was derived with the help of the theory, in the very spirit of Glymour's bootstrapping account and Hempel's satisfaction criterion. Furthermore, the possible observation $\neg Aa.Sa$ would speak against A_1/A_1^* so that the other criterion of bootstrap confirmation – falsification by evidence must be possible – is satisfied, too. Both 'metaphysical' axioms, A_1 and A_1^*, are bootstrap-confirmed by the innocent observation that a person wears a sudra and prays to Ahura-Mazda although they contradict each other. This is unacceptable since apart from A_1 and A_1^*, the two theories are equivalent to each other and still, two outrightly contradictory hypotheses are both confirmed by innocent evidence. True, bootstrap confirmation was intended to model how seemingly neutral evidence gains relevance in a broader theoretical context, but in this particular example, the relevance is far too easy achieved.[25]

An example from the history of science illuminates the problems of bootstrapping in a similar vein. Assume that we would like to test Kepler's Third Law – the quotient of the square of a planet's period and the third power of the average distance to the sun is (roughly) the same for each planet. Call this constant k. Unfortunately, we only have measurements of a single planet which enable us to calculate his orbit and his average distance to the sun. On the other hand, these data are quite diverse, i.e. those quantities could in principle be computed in various ways. But of course, for confirming Kepler's law, we require data from at least two different planets. Now, let $k(a) := T^2(a)/r^3(a)$ be Kepler's constant, calculated for the planet a. Kepler's Third Law now demands that $A_3 = \forall x, y : k(x) = k(y)$, quantified over all planets. This just asserts that the Kepler constant k is the same for each planet. Let $O_i(x)$ denote the i-th observation data of planet x which enables us to calculate his Kepler constant by means of the function f. (Of course, f is the same for each planet.) Then, the auxiliary assumptions can be written as

$$A_1 = \forall x : k(x) = f(O_1(x))$$
$$A_2 = \forall x : k(x) = f(O_2(x))$$

which just means that each relevant measurement O_1 or O_2 opens the way for calculating k. – Assume that we observe planet Mars twice, i.e. we get as our input $O_1(\text{Mars})$ and $O_2(\text{Mars})$. This cannot confirm Kepler's Third Law – for that we would need at least the data of two planets. But the entire theory, $T = \{A_1, A_2, A_3\}$ implies the claim $H' = \forall x, y : k(x) = k(y) = O_1(y)$ which

[25] See Gemes 2006b, 356.

can thus be used in computing the k's. This allows us to bootstrap-confirm Kepler's Third Law by the observations of a single planet: O_1 delivers the Mars data from which we infer to the Kepler constant k of Venus by means of H'. Moreover, O_2 delivers directly the Kepler constant of Mars so that a direct and allegedly meaningful comparison becomes possible.[26] The problem of bootstrapping in both cases, the everyday and the scientific case, can be formulated thus: The dependency between the hypothesis under test and the auxiliary hypotheses H (first example) and H' (second example) turns out to be viciously circular although the hypothesis is put at risk. Therefore it cannot yield a sound case of confirmation.

This objection put us into a dilemma. If we make the conditions for bootstrap confirmation too restrictive, we cannot reconstruct many cases of scientific confirmation. But if we make them too permissive, we obtain cases of spurious confirmation, as seen above. The gullibility objection could be accommodated by noting that bootstrapping models the *coherence* between a piece of evidence and a system of scientific sentences. The above example would then lose much of its pull. But although modeling coherence is certainly important, we are primarily interested in *confirmation.* A restriction of bootstrapping to a formal model of scientific coherence and dismissal of the model of confirmation would abandon the reconstruction of the most interesting cases of confirmation in science – and that was the main idea of bootstrap confirmation. Another reply consists in subtracting the axiom under test from the auxiliary hypotheses which are used in the computations. For instance, we could relativize bootstrap confirmation to a particular axiomatization $\{A_1, \ldots, A_n\}$ and simply demand that the confirmation of any A_i must not rely on A_i itself, but only on $Cn(A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n)$. The problem of gullibility would then vanish. This restriction might, however, invalidate bootstrap arguments in science which use all parts of a theory in confirming an axiom of a theory. Furthermore it is not clear in which way the content of a hypothesis H should be separated from the content of a theory T, so much the more as the notion of a natural axiomatization is notoriously contentious. Hence the gullibility problem stands unscathed. What

[26] Things are actually not that easy since several formal requirements of bootstrapping are violated when using H' in the computation of the Kepler constant. However, in his 1990 Christensen has shown how to circumvent those technical problem by using a suitable modification of H' which maintains the counterintuitive character of the example (Christensen 1990, 651-54).

T:	$A_1 = \forall x Dx \rightarrow Vx$	T':	$A'_1 = \forall x Dx \rightarrow Vx$
	$A_2 = \forall x Dx \rightarrow Bx$		$A'_2 = \forall x Dx \rightarrow (Vx \equiv Bx)$

Table 3.2: Two equivalent theories.

about Christensen's second objection?

Christensen's second charge is directed against the failure of bootstrapping to separate plausible and unacceptable cases of confirmation and to account for evidential relevance. Not only that bootstrap confirmation cannot be sufficient for theory confirmation, it is not even necessary. This criticism goes back to the observation that the way a theory is axiomatized plays a role for bootstrap confirmation. Let us consider two equivalent theories, namely those of table 3.2.

The predicate D is interpreted as the presence of a certain disease, V as the presence of a certain virus and B as a the presence of a certain antibody. Although the deductive closure of both theories is the same ($Cn(T) = Cn(T')$), it seems to make a difference which of the axiomatizations is used and to which axioms we ascribe a lawlike character. The axioms of the first theory claim that a patient with disease D will have virus V and antibody B whereas the axioms of the second theory tell a different story: According to the first axiom, a patient with disease D will have virus V and according to the second axiom, D-patients have virus V if and only if antibody B is present. There is a direct link between antibody B and virus V in theory T'. This motivates observing antibody B *as evidence for the presence of virus V in D-patients.* Such a connection between B and V is missing in T: If $A_1 = A'_1$ is to be confirmed, A'_2 establishes a link between the observation of antibodies and the hypothesis under scrutiny whereas A_2 does not render such an evidence (observation of B-antibodies) relevant. Although both theories have the same set of logical consequences, they seem to mirror different regularities in nature, and a theory of confirmation is supposed to mirror that, or so Christensen argues.[27] Glymour's original bootstrap confirmation fails to make such a distinction: The evidence $E = Ba.Da$ bootstrap-confirms $A_1 = A'_1$ relative to both theories in Glymour's original account and does not bootstrap-confirm $A_1 = A'_1$ relative to either theory in the revised account (Glymour 1983). But the desirable result would be that the evidence confirms A_1 relative to T' (by means of the auxiliary law A'_2)

[27] See Christensen 1983, 479-480, and Christensen 1990, 646-647.

and fails to confirm A_1 relative to T.

This result is obviously awkward. The fatal flaw in bootstrap confirmation seems to consist in the fact that theories are just perceived as deductively closed sets of sentences. Since by the basic principle of bootstrapping, the entire *theory* is used in the confirmation process, the auxiliary hypotheses can be too close to the hypothesis under test. Vicious dependencies between the hypothesis under test and the hypotheses in use emerge. To avoid those problems, we would have to distinguish accidental from lawlike hypotheses in order to separate hypotheses under test from hypotheses in use. As already said, this cannot be done when theories are merely perceived as deductively closed sets of sentences. Therefore a viable modification of bootstrapping has to introduce explicit dependence on the axiomatization of a theory and to separate the hypothesis under test from the hypotheses in use. In the final part of this chapter, we will have a look at attempts to do so.

3.4 Rescuing bootstrapping

The natural way to cure the deficiencies of bootstrapping consists in clearly separating the hypothesis/axiom under test A from the auxiliary hypotheses which are part of the theory. Or, in another vein, the content of the hypothesis has to be separated from the content of the rest of the theory. To this end, two roads can be pursued. First, we might demand that the hypothesis under test A is independent of the 'theoretical background' T', i.e. $T' \not\vdash A$ and $A \not\vdash T'A$. If T is a comprehensive theory that comprises T' as well as A, this means that we have to decompose T' into two parts – an auxiliary part T' and a part A which is to be tested. This proposal is made by Ken Gemes (2006b) in his version of bootstrapping. This step allows him to maintain the satisfaction criterion in bootstrap confirmation although a large number of objections has been made against Hempel's suggestion.[28] The rationale for sticking to the satisfaction condition is Glymour's argument that it is characteristic of many episodes in the history of science, as opposed to hypothetico-deductive confirmation. However, the combination of the bootstrap principle and the satisfaction criterion for elementary confirmation leads into technical problems, as seen above. Only by means of Gemes's stipulation that T' be logically independent of A, we obtain an

[28] See Christensen 1983, 1990 and Horwich 1982.

account of bootstrap confirmation that maintains the satisfaction criterion and is able to deal with Christensen's objections. However, it is not clear how a theory can be 'factorized' into hypothesis under test and hypothesis in use without identifying the theory with a special set of axioms. This is the core principle of the second approach.[29] The motivation for that step was already spelled out – scientists conceive theories not as deductively closed sets of sentences, but as a conglomerate of lawlike statements. Hence, when $T = Cn(A_1, \ldots, A_n)$, we would simply demand that the confirmation of any A_i must not build on A_i itself, but only on $Cn(A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n)$.

It might be objected that relativizing bootstrap confirmation to a particular set of axioms violates the Equivalence Condition which we have unanimously endorsed in the previous chapter. After all, when $\{A_1, \ldots, A_n\}$ and $\{A'_1, \ldots, A'_n\}$ have the same set of logical consequences, it seems odd to treat the corresponding theories in a different way just because different axiomatizations were chosen. From a logical point of view, the two axiomatizations seem to say the same with different words since their deductive closure is identical. Hence, any account of confirmation that does not treat them on a par apparently violates the Equivalence Condition. – I believe, however, that this objection is subtly misguided. The Equivalence Condition attached to single hypotheses that were confirmed by a piece of evidence relative to a set of background assumptions. A scientific theory is more than the sum of its parts and more than just a set of deductively closed sentences – it expresses beliefs about the regularities of nature. Different axiomatizations do not alter the deductive closure of a theory, but they may express different regularities. Therefore the Equivalence Condition is sound with regard to common hypothesis confirmation, but it does not transfer to theory confirmation.

Usually, this second approach additionally replaces the satisfaction criterion by another criterion of confirmation. Douven and Meijs (2006) suggest a Bayesian criterion, but given the focus of the preceding chapter, we would prefer to work in a qualitative framework. Here, the falsificationist criterion has proven to be very valuable in hypothesis confirmation, so it is very natural to to extend its scope to theory confirmation, too. The basic idea is that an axiom of a theory is bootstrap-confirmed if and only it is falsificationally confirmed relative to the other axioms to the theory:

[29] See Christensen 1990, 657-660.

Definition 3.2 *Assume that theory T is the deductive closure of a set of axioms $\{A_1, \ldots, A_n\}$. Evidence E* bootstrap-confirms axiom *A_i, $1 \leq i \leq n$, relative to theory T and background knowledge K if and only if E falsificationally confirms A_i relative to $A_1 \ldots A_{i-1}.A_{i+1} \ldots A_n.K$.*

In other words: When we test an axiom, we take the rest of the theory as additional background knowledge and use it in deducing E from A_i (and $\neg A_i$ from $\neg E$). Hence, evidence that is prima facie not relevant for A_i might gain evidential relevance in the light of a broader theoretical context T. For the classical cases of bootstrapping, the background knowledge K is assumed to be tautologous.

Definition 3.2 extends the falsificationist confirmation of single hypotheses to the confirmation of hypotheses which are part of a theoretical framework. Now we transfer the confirmation relation to whole theories: A theory is bootstrap-confirmed if and only if every axiom of T is bootstrap-confirmed.

Definition 3.3 *Assume that theory T is the deductive closure of a set of axioms $\{A_1, \ldots, A_n\}$. Evidence E* bootstrap-confirms theory *T relative to background knowledge K if and only if E bootstrap-confirms every axiom of T relative to theory T and background knowledge K.*

So bootstrap confirmation amounts to bootstrap confirmation of any axiom of the theory relative to the other axioms of the theory. Hence, definition 3.3 is relative to a particular axiomatization that is supposed to capture the 'natural' regularities in T, as announced before. The precise conditions for bootstrap confirmation of a theory are thus: first, E has to be a content part of $T.K$ because for any axiom A_i, E has to be a content part of $A_i.A_1 \ldots A_{i-1}.A_{i+1} \ldots A_n.K$. This is equivalent to $T.K \models_{cp} E$. That condition accounts for the prediction/observation character of much scientific activity, in the line of the H-D approach. Note that this condition is completely independent of the particular axiomatization employed. Second, any axiom of T has to be falsifiable by the evidence if the other axioms are held fixed. For instance, it is necessary that $\neg E.K.A_2 \ldots A_n \models_{cp} \neg(A_1)_{|dom(E)}.K.A_2 \ldots A_n$, and similarly for all other axioms. The latter condition ensures that the evidence does not bootstrap-confirm theories that contain utterly irrelevant axioms. Every axiom that is bootstrap-confirmed is open to falsification through the evidence. This does not mean that each axiom of a theory is independently testable, i.e. independent of the theoretical background. But this must not trouble us because, after all, bootstrap confirmation is not in-

T:	$A_1 = \forall x : Dx \to Vx$	T':	$A'_1 = \forall x : Dx \to Vx$
	$A_2 = \forall x : Dx \to Bx$		$A'_2 = \forall x : Dx \to (Vx \equiv Bx)$

Table 3.3: Two equivalent theories.

tended as confirmation with regard to *independent* background assumptions. Recall that the motivation of bootstrapping consisted in finding a model for theory confirmation in the absence of external, theory-independent auxiliaries. So bootstrapping describes how a piece of evidence coherently fits into a theoretical picture and how it can act on all relevant parts and axioms of a theory.[30] Hence, definition 3.3 successfully captures the spirit of bootstrapping. Now, let us see how this revised model of bootstrapping deals with Christensen's objections.

David Christensen's main criticism of Glymour's bootstrap confirmation was the lack of an account of *evidential relevance.* Remember the virus/disease case: The problem is that there is a 'natural' connection between B and V in T' which is missing in T. A'_2 establishes a link between the observation of antibodies and the hypothesis under scrutiny whereas A_2 does not render such an evidence (observation of B-antibodies) relevant. Glymour's original bootstrap confirmation fails to make a distinction with regard to the evidential relevance of B-observations, but the falsificationist account of bootstrapping solves the problem: $E = Ba$ bootstrap-confirms $A_1 = A'_1$ relative to $K = Da$ and T', but not relative to $K = Da$ and T. The prediction criterion is fulfilled for both theories, but only T' fulfils the second criterion of definition 3.3 since

$$A'_2.\neg E.K = \neg Ba.Da.\neg Va \models_{cp} Da.\neg Va = \neg A'_{1|\{a\}},$$

but on the other hand,

$$A_2.\neg E.K = \neg Ba.Da \not\models_{cp} Da.\neg Va = \neg A_{1|\{a\}}$$

In the new account, the particular axiomatization of a theory is allowed to reflect nomological relations between the quantities of a theory.[31]

[30] See Meijs (2005, 133-34, 137-40, 162).

[31] The revised bootstrap account in Gemes 2006b comes to a similar result, see Gemes 2006b, 358-359. Unlike me, Gemes sticks to the satisfaction condition, so his version of bootstrapping is closer to Glymour's original work than my proposal.

T:	$A_1 = \forall x Lx \equiv Zx$	$K = Sa$
	$A_2 = \forall x Sx \rightarrow Zx$	$E = Pa$
	$A_3 = \forall x Px \equiv Zx$	

Table 3.4: The 'Zoroastrian Theory'.

With the relevance problem resolved, Christensen's gullibility argument vanishes, too. Christensen interprets predicate L as having eternal life, Z as being a Zoroastrian, S as wearing a sudra and P as praying to Ahura-Mazda (the God of Zoroastrians). In Glymour's original account, A_1 ('all and only Zoroastrians have eternal life') can be confirmed relative to the other axioms of the theory by the observation of a sudra-wearing man praying to Ahura-Mazda. This is obvious nonsense. However, if we apply the novel, falsificationist bootstrap criterion, $E = Pa$ does not confirm A_1 relative to T and $K = Sa$: No observation of sudra-wearing or praying men can ever confirm A_1 since predicate L occurs only in A_1.[32] As long as L denotes an unobservable property, the internal structure of T rules out a confirmation of A_1 since no evidence could ever falsify A_1, even with the help of the other axioms.[33] This shows that Christensen's objections can be satisfactorily answered when bootstrap confirmation is modified in a falsificationist way and relativized to a peculiar set of axioms, at it was proposed by Christensen himself in his 1990.

3.5 Summary

This chapter has discussed how hypothesis confirmation relates to and can be transferred to theory confirmation. As a response to Kuhn's point about the theory-ladenness of observation, it became necessary to develop an account of confirmation that describes how it is possible to confirm theories as a whole, without recourse to external background knowledge. Furthermore the holistic challenge triggered the need for an account of confirmation where entire theories figure as the background against which a hypothesis is tested. We would also like to build a model that describes under which cir-

[32] Technically spoken, there is no way $\neg A_{1|dom(E)}$ could be a content part of $\neg E.K.A_2.A_3$. In the particular case, $\neg E.K.A_2.A_3$ is even contradictory, but the argument in the main text is more general.

[33] Recall that $Lx \equiv Px$ is part of (the deductive closure of) T, but since it is not included in the list of axioms, it is not an admissible auxiliary hypothesis.

cumstances cases of hypothesis confirmation add up to the confirmation of an entire theory. First, such an account could model famous cases of theory confirmation in the history of science. Second, it could describe how evidence is able to prefer an entire theory over a competitor without presupposing theory-independent background knowledge. The main attempt in this direction is Clark Glymour's work on bootstrap confirmation. Glymour's original proposal suffers under severe technical problems so that modifications or re-interpretations have to be made. Either one might adopt a more modest perspective of Glymour's original bootstrapping – in the sense that it models *coherence* between theory and evidence. Or the model of theory confirmation has to single out a particular set of axioms of the theory. Pursuing this road, I have suggested to replace the elementary confirmation predicate in the definition of bootstrapping by the falsificationist criterion from the previous chapter, yielding a viable account of *theory confirmation.* This is not the uniquely feasible way, as Ken Gemes's alternative approach shows, but certainly it is a promising approach to rebut principal arguments for the impossibility of bootstrap confirmation.

Chapter 4

Varieties of Bayesianism

The two preceding chapters were concerned with qualitative confirmation: the question how to define an adequate confirmation relation between sentences of first-order logic. We have encountered two major traditions – confirmation by instances and hypothetico-deductive confirmation and examined their respective virtues and vices. Finally, I have proposed the falsificationist confirmation (FC) as the most adequate confirmation criterion: it is a refined hypothetico-deductive account which respects that typical cases of confirmation are generated by instances of a hypothesis. Falsificationist confirmation can be applied to the more general problem of theory confirmation, too.

Beginning with this chapter, the book will revolve around the question *to which degree* a hypothesis is confirmed by a piece of evidence. A purely qualitative analysis cannot answer this question, and quantitative considerations come into play. For instance, a large part of empirical science uses statistical tools which rely on the probability calculus. Probabilistic data are used to decide between competing hypotheses and to measure which support they lend to a particular hypothesis. The trend towards quantification and probabilification is well-known from the natural sciences and has most recently reached the social sciences. Therefore, there is a strong demand of a philosophy of inductive inference on the basis of probabilistic data. Scientists and policy-makers alike are interested in the degree to which a hypothesis is confirmed, and a solution to the Duhem-Quine problem equally requires a quantitative approach. This chapter and the subsequent one are devoted to the most venerable quantitative theory of confirmation – Bayesian confirmation theory. The Bayesian approach has strong similarity to research in philosophical logic, especially to probabilistic and inductive logic (e.g.

Hailperin 1996). This agrees with the logical positivists' idea of confirmation theory as the *logic of inductive inference.* In qualitative confirmation theory, Carl Hempel took this position as we have seen in the second chapter. In the quantitative tradition, this idea is usually associated with the works of Rudolf Carnap (1950). We will examine in how far Bayesian confirmation theory is able to provide a logic of inductive inference.

The basic idea of Bayesian confirmation can be expressed thus: confirmation consists in the *increase in credibility* of a hypothesis. Before seeing the evidence, we have a certain (rational) degree of belief or credence in the hypothesis at stake. As rational agents, we adapt our degree of belief to the new information after seeing the evidence, in other words, we learn from experience. Evidence that renders the hypothesis more credible is said to support or to confirm the hypothesis, evidence that makes it less credible is said to undermine it. That is, in a nutshell, the main idea of Bayesian confirmation.

Making this idea more explicit requires, of course, the development of appropriate technical tools and raises conceptual problems. We have to ask how degrees of belief can be quantified, elicited and measured and what makes them rational. More precisely, we have to find out the mathematical constraints which rational degrees of belief should satisfy. This is the task of finding a *logic* of partial rational belief, in the same way that deductive logic provides a logic of full rational belief. Finally, we will try to establish a connection to probabilities and the probability calculus as a means of quantifying rational credences. The mathematical theory of probability is often believed to give a *logic* of rational degrees of belief, in the sense that rational agents whose degrees of belief violate the axioms of probability are not fully rational. This can again be compared to deductive logic: agents whose beliefs violate the logical axioms cannot have a consistent system of belief. Probability generalizes this idea to *partial beliefs.* Thus, we have to find an argument that rational degrees of belief should conform to the axioms of probability. Then, the probability calculus would offer a mathematical tool for calculating with degrees of belief and measuring the difference between prior and posterior degrees of belief, as the foundation of a theory of confirmation.

All these questions will be the subject of the present chapter. Subsequently, the results are used to explicate the notion of confirmation with the help of rational degrees of belief. Thus, the chapter does not only lay necessary foundations for the subsequent chapters, but it gives a rough survey on 'subjective probability', too. The reader who is familiar with this subject may therefore skip this chapter and directly proceed to chapter 5. Those who do not feel confident about the connection between degrees of belief and probability and those who are genuinely interested in learning more about subjective analyses of probability, are, however, advised to continue in the text.

4.1 Rational credences

Before explicating confirmation with the help of rational credences, we have to know what rational credences (or degrees of belief) are and how they behave. We follow the classical traditions and take *propositions* as the object of degrees of belief. First, we have to fix a scale for degrees of belief in a certain proposition, and by convention, they usually range between zero (minimal degree of belief) and one (maximal degree of belief), exhausting the set of real (or at least rational) numbers in between.[1] Second, we have to ask how degrees of belief can be elicited and even more fundamentally, what it *means* for an agent to have degree of belief x in proposition A. Credences are communicated by means of (oral or written) utterances, and it is hard to see whether someone who says that he is 'convinced that A is the case' commits himself to any particular numerical degree of belief. If we were to ask this agent 'But to which degree are you convinced? 0.7 or rather 0.8?', such a question would be incomprehensible because the agent would not know what it means to have degree of belief 0.7. She strongly believes that A is the case, but would not see any sense in assigning a particular number to her belief.

As stated above, we are interested in *rational* degrees of belief. The conception of rationality which we utilize is the standard, economical one – irrational degrees of belief would cost us money. Where in real life do degrees of belief have economic significance? When we engage in ordinary action, we are guided by subjective expectations, as Frank Ramsey [1926] (1978) makes clear:

> "[...] all our lives we are in sense betting. Whenever we go to the

[1] Of course, one could also imagine degrees of belief with a completely different range (see Spohn 1990), but the unit interval has some mathematical benefits, and furthermore, standard transformations can be used to map different ranges for credences onto each other.

> station we are betting that a train will really run, and if we had not a sufficient degree of belief in this we should decline the bet and stay at home."[2]

Although our ordinary life is guided by degrees of belief and subjective expectations, the above example does not quantify them. The most pervasive example for quantification are probably transactions on financial markets – stocks, certificates and options are traded and evaluated according to the degrees of belief that they will rise or fall. Someone who has a strong degree of belief that a (European call) option will become worthless evaluates the option's fair price differently than someone who is convinced that the underlying asset will strongly increase the value. For this reasons, options are often called bets on the future. We encounter such bets outside the financial markets, too: People bet on the result of Bundesliga[3] matches, on the future European football champion, and so on. Let's forget at the moment about risk aversion and just presume that those bettors are expected utility maximizers. Obviously, someone who is convinced that Germany will become European football champion will be ready to accept lower betting odds than someone who believes that Germany is just one of several teams which have a good chance to win the cup. Hence, there is a close connection between degrees of belief on the one side and betting behavior on the other side. Therefore, we would like to express degrees of belief as judgments about fair betting odds and map both quantities to each other. Under a *bet* on the event A we understand a triple $\langle A|x|y\rangle$ where x and y are positive real numbers: The bookie pays the bettor y Euro if A occurs and the bettor pays the bookie x Euro if A does not occur. x is the bettor's *stake*, and the ratio $(x+y)/x$ is called the *betting odds* on A, indicating the bettor's total gain (including the stake) for a successful 1 Euro bet.[4] An agent believes betting odds to be fair if they offer no advantage to either side, i.e. if to the agent's mind, the bookie and the bettor have the same expected utility. Consequently, an agent believes proposition A to degree p if he believes the associated bet to

[2] Ramsey [1926] 1978, 85.

[3] The 'Bundesliga' is the first German football division.

[4] Diverging from the convention in most of the philosophical literature, I use *decimal odds* which are popular in Continental Europe instead of *fractional odds* that prevail in the United Kingdom. Fractional odds give the net return the bettor gets for a successful 1 Euro bet, decimal odds give the total return for a successful 1 Euro bet, including the initial stake. Decimal odds are easier to calculate with, e.g. when combining bets on several events into a single bet, the total odds are just the product of the odds of the components.

be fair. Numerical degrees of beliefs express judgments about the fairness of bets.

It seems that betting odds offer a convenient way to express degrees of belief. When we believe a sentence to be true low betting odds seem to be fair whereas, when we believe a sentence to be false, we would ber ready to grant high betting odds. So there is an intuitive connection between betting odds and degrees of belief. Now it could be objected that an explication of degrees of belief that relies on judgments about the fairness of bets is much too close to the original concept (degree of belief) and therefore not very illuminating. Furthermore, the notion of fairness used in this 'definition' of degrees of belief is problematic due to the vagueness inherent in that concept. We might not be able to reveal the degree of beliefs of persons whose concept of fairness deviates from our concept. Therefore we need a standard definition of fairness: we might ask the agent to imagine that she is going to be either the gambler or the bookie, but only after designing the bets we will tell her which side she is going to play. This technique resembles John Rawls's (1971) 'veil of ignorance' for disclosing judgments about the fair distribution of goods in a society. This operationalist definition of fairness gives an argument for a fully dispositional, behaviorist definition of degrees of belief (e.g. by De Finetti (1937)): agent S's degree of belief in proposition A is equal to p if and only if p utility units is the price at which S would sell or buy a bet on A that pays 1 utility if A occurs.[5] In other words, p is the degree of belief where an agent would be indifferent between the two sides of the bet if she were forced to choose. Or, put even another way, p is the degree of belief she would submit under the veil of ignorance (whether she will be the bettor or the bookie). These operationalist definitions match, of course, the program of logical positivism. But on the other hand, they entail various problems, e.g. the agent may draw some extra utility from being a particular side of the bet.[6] It is hard to define the ideal circumstances and the ideal agent where an operationalist definition would apply. In particular, believing betting odds to be fair does not mean that we would *actually* be willing to take any side in the bet. We might be so risk-averse that we do not engage in any bets for fear of losing the stake, even if we believe that the bet is quite advantageous for us. This threatens the operationalist explication of degrees of belief as hypothetical betting behavior – many people will for principled

[5] See Hájek (2007).

[6] See Mellor 2005, 65-66.

reasons never engage in bets, thus they will never have dispositions to accept certain bets. Therefore Howson and Urbach (1993) write:

> "To believe odds to be fair is to make an intellectual judgment, not [...] to possess a disposition to accept particular bets when they are offered."[7]

To say that a person has degree of belief x that event A will happen does thus not mean that he would exhibit any particular betting behavior – it just says something about which kind of bets he considers advantageous for a particular side. Even someone who would never accept any bets seems to be able to make judgments about the fairness of bets and to entertain numerical degrees of belief. It is more convenient and less problematic to conceive credences as judgments about the fairness of betting odds than as dispositions to accept and to reject certain bets.

But wait a moment – didn't we say that numerical degrees of belief are normalized to the unit interval $[0, 1]$? Betting odds, however, live in the interval $[1, \infty[$. We showed that degrees of belief can be expressed by judgments about fair betting odds, but they do not directly correspond to them. Hence, we have to build an isomorphic map between the two intervals. The higher the fair betting odds for an event A, the lower the agent's credence in A. The intuitive idea of fixing credences between zero and one is that $p \in [0, 1]$ expresses the credence that A actually occurs in fraction p of all possible courses of events. Intuitively, it is clear what we mean by such a 'definition' – it explains the choice of the unit interval as the range for degrees of belief, and in the next subsection, the benefits of this convention will be seen. Nonetheless, it is a very abstract and vague idea and not suitable for giving empirical *meaning* to credences. How can we establish the connection between degrees of beliefs and betting odds? If agent S believes A to occur in $(100 * p)\%$ of all possible cases, he will consider the bet $\langle A|x|y\rangle$ to be fair if and only if the zero-sum condition is satisfied:

$$p\,y + (1-p)(-x) = 0. \tag{4.1}$$

The only value in $[0, 1]$ that solves equation (4.1) for $x, y > 0$ is $p = x/(x+y)$. Hence, we can determine the degree of belief in an event A, understood as the putative fraction of successful bets to total bets on A, from the fair

[7] Howson and Urbach 1993, 57.

betting odds by taking the inverse. Similarly, if someone entertains credence p in A, he commits himself to the fair betting odds $1/p$. Hence, there is an isomorphism between credences and betting odds. High credences yield low betting odds and low credences yield high betting odds, as we know it all from betting on sports events. Someone who assigns degree of belief p in a proposition is, by the above isomorphism to betting odds, able to check whether his degree of belief really corresponds to a bet which she would consider fair.

So far we have introduced degrees of belief and explained the connection between betting odds and degrees of belief. Typically, we have degrees of belief in more than one proposition, e.g. there are two propositions A and B in which we have definite degrees of belief. Then, this should put constraints on the degrees of belief in truth-functional compounds of A and B: for instance, the degree of belief in $A.B$ should be lower than each of the degrees of belief in A and B. (The conditions for winning a bet on $A.B$ are harder than the conditions for winning a bet on either A or B so that a bet on $A.B$ should have higher betting odds. Since higher betting odds entail lower degrees of belief, $A.B$ should be believed to a lower degree than A and B, independent of the meaning of the propositions.) To elaborate that idea, we have to define when a set of propositions is closed under the usual truth-functional operators (conjunction, disjunction, negation, etc.).[8] Such a set is called a *field* or *algebra*.

Definition 4.1 *A* field *of propositions $\mathcal{A}$ is a set of propositions so that*

- *any tautology is in $\mathcal{A}$*
- $A \in \mathcal{A} \Rightarrow \neg A \in \mathcal{A}$
- *If A_1, $A_2 \in \mathcal{A}$, then also $A_1 \vee A_2 \in \mathcal{A}$.*

For instance, assume that we have degrees of belief in the propositions A, B and C, each of them expressing assertions about the world. Then a field of propositions that contains A, B and C also contains all truth-functional compounds of A, B and C, e.g. $A.\neg B$, $A \vee B \vee C$, etc. On the other hand, a

[8] This entails that tautologies and contradictions are also part of that set. From any proposition A, we can construct a tautologies and contradictions by means of $A \vee \neg A$ and $A.\neg A$.

field of propositions that contains A, B and C does not necessarily contain proposition D if D is completely independent of A, B and C.

For technical reasons, the notion of a field is often amended to a sigma-field (σ-field) which satisfies an additional condition, namely closure under countable disjunction:

Definition 4.2 *A* σ-field *of propositions* $\mathcal{A}$ *is a field of propositions so that*

- *If* $A_n \in \mathcal{A}$ *for all* $n \in \mathbb{N}$*, then also* $\bigvee_{n \in \mathbb{N}} A_n \in \mathcal{A}$.

Hence, a σ-field of propositions is closed under logical negation and *countable disjunction* (and thus also under countable conjunction). Since the countable additivity condition plays a major role in statistical applications of quantitative confirmation theory, we will from now on work with σ-fields instead of mere fields.[9]

4.2 Probability and the Dutch Book Arguments

Assume that we have degrees of belief in all propositions that belong to a sigma-field. Our system of beliefs is thus closed under truth-functional combination. We are now concerned with the following questions: how do the degrees of belief have to cohere with each other in order to form a consistent system of partial beliefs and to count as *rational* credences? An argument for a very simple constraint has already been given above: the credences in

[9] It might be objected that most countable disjunctions of standard propositions cannot be written down in closed form. Therefore it is hard to intuitively make sense of the condition that the σ-field be closed under countable disjunction – finite disjunction seems to do the job, too. Two points can be replied. First, there are some very important countable disjunctions of propositions – namely existential claims on a countable domain. The sentence 'There is a natural number that is the square root of 49' is the disjunction of the sentences '1 is the square root of 49', '2 is the square root of 49', etc. Sometimes there are very elegant linguistic ways to express countable disjunctions of standard propositions. Second, the problem that some of those countable disjunctions do not admit a closed form can just be regarded as a problem of minor importance. Neither do countable disjunctions of *sets* always admit a closed form even if any of the constituting sets has a closed form. Nevertheless, such sets are very important in set theory. Moreover, we are interested in applying our theory of credences to hypotheses about the empirical world. Such hypotheses may be very strange disjunctions of 'elementary' hypotheses, and there is no reason to rule out such hypotheses in advance.

the conjunction of two propositions should not exceed the minimum of the credences in each proposition. But there is more to be said as we will soon see. Let $P(A) \in [0,1]$ be the degree of belief in proposition A. First, all logical truths A should be assigned maximal degree of belief $P(A) = 1$ because bets on them cannot lose. So, any betting odds that deviate from 1 would be unfair to one of the two sides.[10] In a similar vein, all contradictions B should have minimal degree of belief $P(B) = 0$ – such bets can never be won so that only infinitely large betting odds would be fair. Second, assume that A is a contingent proposition. There is an alternative description for a standard bet $\langle A|x|y\rangle$, namely that the bookie bets on $\neg A$ and that the bettor takes the bookie's position. After all, the bettor's net gain in the case of success corresponds to the bookie's stake and the bettor's stake corresponds to the bookie's net gain in the case of failure. So, the betting odds on A also fix the betting odds on $\neg A$: If the bettor makes a wager on $\langle A|x|y\rangle$, then the bookie effectively makes a wager on $\langle \neg A|y|x\rangle$. Hence, if the fair betting odds on A are $(x+y)/x$, the fair betting odds on $\neg A$ are $x + y/y$. For degrees of belief, this entails that $P(A) = x/x + y$ whereas $P(\neg A) = y/x + y$. In other words,

$$\begin{aligned} P(\neg A) + P(A) &= \frac{x}{x+y} + \frac{y}{x+y} \\ P(\neg A) &= 1 - P(A). \end{aligned} \tag{4.2}$$

Hence, the degrees of belief in a sentence and its negation add up to 1 so that the more credible a sentence, the less credible its negation. This is arguably a nice property of degrees of belief and motivates once more the choice of the unit interval as the range of degrees of belief. Finally, it makes some sense to postulate that for two mutually exclusive propositions, the credence that one of them will come true exceeds the credence in each of the two propositions. More precisely, it is plausible to ask for *additivity*: for two propositions A_1 and A_2 with $A_1 \models \neg A_2$, the degree of belief in $A_1 \vee A_2$ is equal to the sum of the degrees of belief in A_1 and A_2 – more on this later.

Thus, we have already spot some conditions which a consistent system of rational credences should satisfy. The mathematical tool which is often believed to describe the logic of rational credences and partial belief is the probability calculus. This may sound a little bit far-fetched, but we can spot

[10] Clearly, this makes the idealizing assumption that agents are logically omniscient; otherwise they would be unable to judge the unfairness of such a bet.

the connection to degrees of belief in a variety of examples from ordinary speech. People use probabilities in order to qualify the strength of their conviction, i.e. if event A is said to be probable, it is also believed to be subjectively credible. But what precisely is the connection between degrees of belief and probabilities? To answer this question, we should introduce probabilities as mathematical objects. A probability is a measure on a sigma-field that satisfies the following constraints (see Kolmogorov [1933] 1956):

Definition 4.3 *Let $\mathcal{A}$ be a sigma-field of propositions. $P : \mathcal{A} \rightarrow [0, 1]$ is a* probability function *on $\mathcal{A}$ if and only if*

- $P(A) = 1$ *for any tautology A.*
- $P(\neg A) = 1 - P(A)$.
- *σ-additivity: For pairwise mutually exclusive propositions $A_1, A_2, \ldots$ (i.e. $A_i \models \neg A_j$ for all $i \neq j$), $P(\bigvee_{n \in \mathbb{N}} A_n) = \sum_{n=1}^{\infty} P(A_n)$.*

The mathematically educated reader will have noticed that I introduced probabilities different from their introduction in standard textbooks on probability theory. In the mathematical literature, probabilities assign real numbers to subsets of a (measurable) space and not to propositions. However, the set-theoretic operators of complement, intersection and disjunction correspond to the logical operators of negation, conjunction and disjunction so that the definitions are more or less isomorphic. The sentential approach was chosen in this presentation because it suits our intended applications of probability theory better: we are interested in interpreting probabilities as degrees of belief in empirical hypotheses. Then it is natural to define probabilities for sigma-fields that consist of propositions and not of sets.

The three items of the above definition are henceforth called the *axioms of probability.* The first two items of the above definitions implement our motivational remarks for degrees of belief (e.g. equation (4.2)) whereas the third axiom still requires substantiation. More generally, while it looks plausible that degrees of belief conform to the probability axioms, we have not yet given a solid argument for this case. Why does a system of credences have to conform to the probability calculus in order to count as *rational* credences? Put another way, why would it be irrational to deviate from the axioms? The upshot is thus: Our degrees of beliefs determine betting odds for all events in the sigma-field. Now, if one of the probability axioms is violated, then the entire system of odds cannot have been fair – in the sense that it is possible to construct a system of bets that assures a riskless positive gain to the bookie or the bettor. In other words, arbitrage becomes possible. Such a system of bets is called a *Dutch Book*, and the associated theorem – if the axioms are violated, Dutch Books can be made – is called the *Dutch Book theorem.*[11]

Theorem 4.1 *Any function* $P : \mathcal{A} \rightarrow [0, 1]$ *on a* σ*-field* $\mathcal{A}$ *that does not satisfy the axioms of probability induces a system of bets that is vulnerable to a Dutch Book. This means that there is a system of bets on elements of* $\mathcal{A}$ *according to the betting odds given by* P *so that one side in the total system of bets is guaranteed a positive net return.*[12]

Proof: See appendix A.

The Dutch book theorem establishes that obeying the axioms of probability is necessary for having a coherent system of fair betting odds. A system of bets where one side has the opportunity of a riskless gain cannot be fair. This directly transfers to degrees of belief if they are explicated as judgments about fair bets:

> "If anyone's mental condition violated these laws [of the probability calculus], his choice would depend on the precise form in which the options were offered him, which would be absurd."[13]

Thus, probability provides a calculus for the logic of partial belief.[14] The converse Dutch Book Theorem, proved by Kemeny (1955), establishes the counterpart of the Dutch Book theorem – any system of partial beliefs that obeys the axioms of probability is indeed coherent, i.e. immune to Dutch Books. Or in other words, if the axioms of probability are satisfied, no Dutch Books can be construed. On the other hand, the Dutch Book theorem does

[11] For a more detailed discussion, consult e.g. the second chapter of Earman 1992, the third chapter of Howson and Urbach 1993 or the earlier presentation in Kemeny 1955. The converse theorem (if the axioms are not violated, there are no Dutch Books) is mentioned later in the text.

[12] The normative force of Dutch Book arguments, in particular when combined with σ-additivity, is a subject of philosophical debate, but it cannot be discussed here.

[13] Ramsey [1926] 1978, 84.

[14] This does not mean that probability calculus is the only conceivable logic of partial belief – see, for instance, Spohn 1990, 2008.

not put constraints on any specific degree of belief – it just states that if we were to entertain degrees of belief in a set of propositions, our degrees of belief have to respect the logical dependencies between those propositions, on pain of irrationality.

Degrees of belief were explicated by means of judgments about fairness in a hypothetical betting game. A system of partial beliefs in a set of interconnected propositions is rational only if it is internally consistent, i.e. immune to Dutch books. The Dutch Book theorem establishes a connection between degrees of beliefs and probabilities and explains why judgments about probabilities are closely related to credences – they share the same calculus.[15] But we should distinguish two ways of speaking about probabilities – the informal, everyday concept and the mathematically precise definition of probability. So far, we were mainly interested in the connection between degrees of belief and the *mathematical* object probability, in particular because rational degrees of belief respect the probability calculus. But there is also an informal concept, and it has two different interpretations – the subjective and the objective interpretation. If A is said to be probable, this can be understood as an assertion that A is credible in the eyes of the agent. Here, probability is interpreted as degree of belief. On the other hand, a statement of probability often seems to say something about the *objective chance* of an event, independent of all personal degrees of belief. Classically, probabilities are either interpreted as degrees of belief or as objective chances. The objective probability of an event is normally either a property of a real-world processes or the limiting relative frequency of the occurrence of an event in a repeatable trial. Clearly, objective chances do not supervene on degree of beliefs. Physical theories often use the objective interpretation, e.g. when it is claimed that a deuterium atom has a 50% chance to decay within a year. In a more mundane context, a particular die might have an objective probability/tendency of 1/6 to come up with a 'six'. Both examples use

[15] Again, countable additivity for rational credences might be questioned. But statistical applications ask for countable additivity: Assume we have a integer-valued stochastic process, and we are convinced that the process will eventually terminate. Nevertheless, we have absolutely no idea at which point it will terminate. Denying countable additivity would allow us to assign rational credence zero that the process will end at time point n, *for all* $n \in \mathbb{N}$. Now, there is an awkward tension between our conviction that at some point, the process *must* terminate and our conviction that for any moment n, it is impossible that the process terminates there. Intuitively, such a system of credences does not seem to qualify as rational. Countable additivity circumvents that problem.

the objective interpretation which is completely independent of belief states. There is, however, an instructive connection between the subjective and the objective interpretation which we are going to examine in the next section.

4.3 The essence of Bayesianism

The previous section has outlined how degrees of belief can be expressed by fair betting odds in hypothetical games. Furthermore, the Dutch Book theorems have argued that rational credences conform to the axioms of probability. But note that this justification only extends to the consistency of a system of beliefs that are held *simultaneously*! We did not give a principle how degrees of belief should be *changed* in the light of incoming information. Put another way, we have motivated synchronic, but no diachronic rationality constraints on degrees of belief. The crucial idea of Bayesianism is that learning from experience can be expressed by conditional probabilities. The conditional probability of a proposition A given another proposition B is defined as

$$P(A|B) = \frac{P(A.B)}{P(B)}. \tag{4.3}$$

Conditional probabilities (or conditional degrees of belief) correspond to bets on A that are canceled if B does not occur. We can read (4.3) as the fraction of 'cases' where A and B are true, relative to the total number of cases where B is true.[16] In other words, the conditional probability of A given B can be interpreted as the rational credence in A if we believed B to be true. Thus, our rational degree of belief that a proposition H is true in the light of incoming information E or after learning E is the *conditional probability* of H given E. Degrees of beliefs are changed by conditionalizing on incoming information. The principle of **Bayesian Conditionalization** describes how to update degrees of belief in the light of incoming information:

$$P_{\text{new}}(H) = P(H|E). \tag{4.4}$$

(4.4) is also called the principle of strict conditionalization. Applying Bayes's theorem, a simple fact of probability theory, to calculate the conditional

[16] A major approach in the philosophy of mathematics takes conditional probabilities as primitive and derives 'normal' probabilities from conditional probabilities, using an isomorphic axiomatization, see Hájek 2003.

probability yields

$$P_{\text{new}}(H) = P(H|E) = \frac{P(H)\,P(E|H)}{P(E)}. \tag{4.5}$$

Here, we see the connections between Bayesianism and probabilistic inductive logic. The probabilities of some sentences impose constraints on the probability of other sentences, according to the laws of probability. Hailperin (1984, 205) has suggested the following scheme: For sentences $X_1, \ldots X_n, Y$ and $\alpha_1, \ldots, \alpha_n, \beta \in [0,1]$

$$P(X_1) = \alpha_1, \ldots, P(X_n) = \alpha_n \quad \models \quad P(Y) = \beta \tag{4.6}$$

just in the same way that the valuations A = true and $A \rightarrow B$ = true logically imply that A = true.[17] In the case of Bayesian conditionalization that we are interested in, the values of $P(H)$, $P(E)$ and $P(E|H)$ determine the values of $P(H|E)$, the posterior credence in H:

$$P(H) = \alpha_1,\ P(E) = \alpha_2,\ P(E|H) = \alpha_3 \quad \models \quad P(H|E) = \frac{\alpha_1 \alpha_3}{\alpha_2}. \tag{4.7}$$

In other words, the *posterior* degree of belief in H – the degree of belief in H after learning E – is the product of the *prior* degree of belief in H and the likelihood of E given H, divided by the expectedness of the evidence E. Bayesian inference can thus be represented by means of consequence relations in a probabilistic inductive logic that extends deductive logic.[18]

How do we compute the probabilities on the left hand side of (4.7)? The probability of the evidence can be decomposed as

$$P(E) = P(H)\,P(E|H) + P(\neg H)\,P(E|\neg H) \tag{4.8}$$

and thus be traced back to the likelihood of E given H and $\neg H$ as well as the prior probability of H and $\neg H$. We see that Dutch Book arguments are complementary to Bayesian Conditionalization: whereas Dutch Book arguments govern the *statics* of degrees of belief, conditionalization describes the *dynamics* of degrees of belief. Together they are able to give a theory of partial rational belief and learning from experience: Degrees of belief are rational if and only if they emerged from a consistent system of degrees of

[17] The consequence relation in (4.6) is, of course, not deductive entailment, but a *probabilistic* consequence relation.

[18] See also the works of the PROGIC research group, e.g. Haenni et al. 2008.

belief by Bayesian conditionalization. This is the *orthodox Bayesian position* (De Finetti 1937): Subjective credences are rational if and only if they conform to the probability calculus and use Bayesian conditionalization as the updating rule in the face of new evidence.

To give an example of Bayesian Conditionalization at work: Assume that we know that a specific coin is either strongly biased towards heads, with a 3:1 proportion of heads over tails (H), or it is a normal coin with equal tendency to fall heads or tails ($\neg H$). We have equal prior credences in H and $\neg H$ ($P(H) = P(\neg H) = 1/2$). The coin comes up tails (E). If H is true, this evidence has a 1/4 chance of occurring ($P(E|H) = 1/4$) and it has a 1/2 chance of occurring when H is false ($P(E|\neg H) = 1/2$). Then, the probability of this event is

$$\begin{aligned} P(E) &= P(H)\,P(E|H) + P(\neg H)\,P(E|\neg H) \\ &= \frac{1}{2}\frac{1}{4} + \frac{1}{2}\frac{1}{2} = \frac{3}{8}. \end{aligned} \tag{4.9}$$

Hence, in the light of (4.9), the posterior probability of H (the hypothesis that the coin is biased) in the light of the observed evidence becomes

$$\begin{aligned} P_{\text{new}}(H) &= P(H|E) = \frac{P(H)\,P(E|H)}{P(E)} \\ &= \frac{1}{2}\frac{1}{4}\frac{8}{3} = \frac{1}{3}. \end{aligned} \tag{4.10}$$

We see that the credence in H has decreased from 1/2 to 1/3 due to the relatively unexpected event that the coin came up tails. This calculation illustrates the way Bayesians change their degrees of belief in the light of incoming information.

But wait a moment – didn't we conflate the subjective and the objective concept of probability in the above calculation (4.10)? We correctly stated that the *objective chance* of the event 'tails' under H is equal to 1/4. Nevertheless, there is no *logical* point why this should transfer to rational credences, too. So far, we have not clarified whether the probabilities that occur in Bayes's theorem are subjective or objective, but standardly, they are considered to be subjective (since Bayesianism is a theory of belief revision). Then, the above example of Bayesian updating needs more presuppositions: We have only assumed that the objective chance of E under H equals 1/4, but we did not mention any subjective credences in E given H or $\neg H$, in other words, we did not explicitly fix the value of $P(E|H)$ and $P(E|\neg H)$. An orthodox Bayesian might object to equation (4.10) and simply assert that for some reason, he has degree of belief 1/3 that tails will come up if H is true. This assignment is certainly strange, because objective chances then differ from subjective credences. But it violates neither the axioms of probability nor the conditionalization principle, and thus, the charge of irrationality cannot be raised against the orthodox Bayesian.

Nevertheless, we somehow feel the need to calibrate our degrees of belief to objective chances. Whatever those chances denote – limiting relative success frequencies, tendencies of causal processes to yield a particular event or the fraction of a subset of possible worlds – they always guide our subjective expectations about future events. For instance, if a process has an objective chance (e.g. a causal tendency of) 1/4 to yield success, we should adapt our rational credences to that value if we happen to know it. This principle is so intuitive that some authors even believe it to be an analytic statement. This means that our degree of belief in an event should correspond to its objective chance in the external world whenever we know the value of such chances. For instance, if we knew that the objective chance of a 'six' in the toss of a die were 1/6, we should set our rational credence to that very value (since degrees of belief express the fraction of 'successful' and possible courses of events). In this way, our background knowledge about objective chances puts constraints on our rational degrees of belief.

How can we formulate this principle? Assume that a function P on a σ-field $\mathcal{A}$ expresses rational degrees of belief of a subject (and thus satisfies the axioms of probability) and let S be a proposition that asserts that a certain event A occurs at time t. Furthermore, let A_p be the proposition asserting that the objective chance of the occurrence of A at t is equal to p. Now, the rational credence in S given the information A_p is equal to p, or in mathematical form,

$$P(S \mid A_p.K) = p \tag{4.11}$$

where K is any background knowledge that is admissible[19] at point t and compatible with S.[20] This is David Lewis's (1980) *Principal Principle*: If

[19] The notion of admissibility is a technical one and explicated in Lewis 1980. The point is to rule out information that 'interferes' with the proposition A_p. Most standard background information, e.g. historical information about matters of fact prior to t, satisfies the admissibility condition.

[20] This treatment is analogous to Earman (1992, 51-52).

we know that the objective chance of a certain event A (whatever this is) is equal to a number p, we should adapt our credence in the occurrence of A to that very same number. Rational degrees of belief track objective chances. This principle has to be refined, of course, in order to be immune against counterexamples.[21] For the purpose of this book, though, the basic sketch is sufficient. Let's come back to the above example of (4.10). If we maintained credences different from 1/4 in $P(E|H)$ (tails is the result of a toss of the biased coin), we would be irrational. Our credences would imply that we consider betting odds different from 1:4 to be fair for both sides. But since the objective chance of tails is equal to 1/4, such a bet cannot be fair – there will merely be a 25% success rate in the long run. So if we knew that tails had a 1/4 chance of coming up, our rational degrees of beliefs should be fixed to that very number.[22] Hence, the above example for Bayesian Conditionalization only works if we amend the Conditionalization Principle with the Principal Principle. In fact, the Principal Principle is so intuitively plausible and self-evident that most Bayesians accept it. If we combine the axioms of probability with conditionalization and the Principal Principle, we adopt the *personalist Bayesian* position or the *empirically-based subjective Bayesian* position: All degrees of belief that conform to the axioms of probability, emerge from Bayesian conditionalization and take into account empirical constraints (via the Principal Principle) count as rational. Degrees of belief which violate one of these constraints are irrational.[23] In other words, we could say that the personalist Bayesian requires both internal

[21] See Lewis 1980 and, for a comprehensive discussion of different versions of the Principal Principle, Rosenthal 2004.

[22] Quite often, the likelihood of evidence under a statistical hypothesis is fixed. It seems to be part of the *meaning* of H and E in the above example that the conditional probability of E under H is 1/4. To see that in greater detail, note that the sentence $P(E|H) = 1/4$ can be reformulated as

> 'If there is a 3/4 probability that the coin comes up heads (=H), the probability that the coin comes up tails (=E) is 1/4.'

has *analytic character*. Regardless of which interpretation of probability we select – someone who objects to the above sentence seems to have misunderstood the meaning of E and H. Rational credences should not only track objective *physical* chances, but also those chances which are fixed by the meaning of the assumptions (H) on which we conditionalize. As mentioned at the beginning, this frequently occurs when we calculate the likelihood of events under statistical hypotheses.

[23] A representative of this popular view is Colin Howson (2003) as well as most Bayesians that are not explicitly associated with the objectivist or the orthodox position.

validity of her credences (no Dutch book is possible) and external validity (her credences correspond to objective chances).[24]

Still, in the personalist perspective, there is usually a plurality of rational credence functions. Not all researchers in the field find that appealing. Therefore, the rationality constraints on degrees of belief are sometimes further sharpened: there is only a single rational degree of belief. This is the tenet of *objective Bayesianism.*[25] Objective Bayesians go beyond the personalist positions by claiming that only one of the belief functions that is admissible for a personalist is factually a rational belief function.

Before comparing and discussing the various forms of Bayesianism, we should discuss two objections to the standard Bayesian machinery. First, the information on which we update is often not certain and subject to transmission or measurement errors so that it is not completely clear whether E or $\neg E$ has occurred. Hence, it seems unwise to fully conditionalize on that information and to take it for certain. To this end, the conditionalization rule can be modified in the following way, proposed by Richard Jeffrey:

$$P_{\text{new}}(H) = q\, P(H|E.K) + (1-q)\, P(H|\neg E.K), \qquad q \in [0,1]. \tag{4.12}$$

(4.12) takes the weighted average of the posterior probability of H under E and $\neg E$, taking $q \in [0,1]$ as a factor that quantifies the uncertainty whether E has really occurred. This principle is named after his author – **Jeffrey Conditionalization (JC)** –, and in the special case $q = 1$, it coincides with the strict conditionalization principle. Jeffrey Conditionalization generalizes Bayesian Conditionalization and accounts for evidential uncertainty.

Second, Bayesian updating can have some curious consequences. For instance, we are in principle allowed to assign maximal degree of belief to the proposition H that the moon consists of yellow cheese. According to the probability calculus, such extreme beliefs can never be revised since the negation of that proposition has degree of belief zero ($P(\neg H) = 0$), entailing $P(E) = P(E|H) = 1$. It is certainly strange to count someone as a rational agent who dogmatically sticks to his conviction even in the face of strongly undermining evidence. Rational agents should be open-minded towards empirical hypotheses, i.e. no empirical and contingent statement should ever

[24] This point was brought to my attention by Andreas Bartels.

[25] I omit the discussion of interval-valued probabilities. The main topics of the book are confirmation theory, inductive inference and statistical reasoning, and the debate about rationality constraints on degrees of belief should not distract too much from the red line.

receive degree of belief 1 or 0 since it might always be the case that we are mistaken in our beliefs about the external world. Therefore, we might amend the various Bayesian positions by a principle of *regularity* or *non-dogmatism* (R). Rational agents should in principle be responsive to evidence:

> (R) In a system of rational degrees of belief, only logical falsehoods (contradictions) are assigned probability 0 and only logical truths (tautologies) are assigned probability 1.

4.4 Varieties of Bayesianism

So far, we have encountered roughly three different Bayesian positions. First, the subjective or orthodox Bayesian position where all degrees of belief that respect the axioms of probability and Bayesian conditionalization count as rational. Second, the personalist Bayesian position which supplements the first view by the condition that rational degrees of belief have to track objective chances, i.e. the Principal Principle. Finally, the objective Bayesian position asserts that there is only one rational degree of belief for each proposition and that rational belief requires more than respecting empirically given constraints.

The aim of this section consists in illuminating the virtues and vices of these three basic forms of Bayesianism (which may, or may not, be amended with constraints such as regularity). It is clear that the personalist position is more attractive than the orthodox one, due to the intuitive plausibility of the Principal Principle. It is less clear, however, whether the personalist position is superior to the objective Bayesian view, too. This contrast will be the main focus of the section.

The main principle of a particularly appealing form of objective Bayesianism is *entropy maximization* and has been proposed by E. T. Jaynes (1957, 1968).[26] Assume that we have a finite space of possible events or outcomes $\Omega = \{x_1, x_2, \ldots, x_n\}$. This set Ω generates a (set-theoretic) sigma-field $\sigma(\Omega)$ that comprises all disjunctions, intersections and complements of subsets of Ω. We can also interpret the elements of Ω as complete descriptions of matters of fact in a system. For instance, suppose that three propositions A, B and C describe elementary matters of fact in a toy world $\mathcal{W}$ (e.g. $\mathcal{W}$ is

[26] See also Shannon and Weaver 1949 and Uffink 1996. The presentation in this book follows Williamson 2007.

a system of three independent electric bulbs which can be switched on and switched off). Then, $\Omega = \{T, F\}^{A,B,C}$, e.g. the elements of Ω such as $A.B.C$ or $A.\neg B.\neg C$ describe bulbs are switched on and which are switched off. Consequently, the elements of $\sigma(\Omega)$ exhaust all kind of first-order assertions about matters of fact in $\mathcal{W}$, e.g. $A.\neg B$ or C.

Let $\mathcal{P}_K$ be the set of probability distributions on the measurable space[27] $(\Omega, \sigma(\Omega))$ that are compatible with the agent's background knowledge K. Objective Bayesians now single out the element of $\mathcal{P}_K$ which minimizes the distance to the uniform distribution on Ω, i.e. the distribution that assigns equal weight to each element of Ω (i.e. $P(x_i) = 1/n$ for all $i \leq n$). In other words, objective Bayesians recommend to choose the most equivocating distribution among all distributions that are compatible with the agent's background knowledge.

How is the distance to the uniform distribution measured? Usually, the *cross-entropy* (Kullback and Leibler 1951) of the two distributions is utilized. It figures under the names Kullback-Leibler (K-L) discrepancy, K-L divergence, K-L information or relative entropy, too. The standard motivation for using K-L divergence stems from coding theory when a string of symbols from Ω has to be transmitted by means of 0-1-electric signals. To save time and energy, this string is compressed during the transmission, i.e. frequently occurring symbols are assigned short sequences of zeros and ones whereas rarely occurring symbols are coded as longer sequences of zeros and ones. Now, it may happen that we do not know the precise distribution f of the symbols x_i in the entire string (i.e. the relative frequencies of the x_i). Instead, we use an approximating distribution g. Then, the expected loss of efficiency in the transmission that emerges by using g instead of the optimal distribution f is given by the *cross-entropy*

$$H(f, g) = \sum_{i=1}^{n} f(x_i) \log \left(\frac{f(x_i)}{g(x_i)} \right) \tag{4.13}$$

where the sum is taken over the entire sample space Ω.[28] $\log[f(x_i)/g(x_i)]$ measures the information loss between f and g for every data point x_i, and averaging the loss according to $f(x_i)$ (the relative frequency of a symbol x_i) yields the expected information loss. Note, however, that this function is

[27] A measurable space is a space plus a sigma-field on that space.

[28] The definition easily generalizes to the continuous case, too – the sum is replaced by an integral.

not symmetric, hence it is no 'distance' in the proper sense of the word. The terminus 'discrepancy' suits better.[29] Cross-entropy is a mathematically tractable quantity and enormously significant in mathematical statistics, information theory and even statistical mechanics (Boltzmann entropy). For these reasons – and because there exists a clear, crisp and application-oriented motivation – it is the most prominent measure of discrepancy between two probability distributions, despite the lack of symmetry.[30]

The cross-entropy between an element P from $\mathcal{P}_K$ and the uniform distribution U can then be written as

$$\begin{aligned} H(P,U) &= \sum_{i=1}^{n} P(x_i) \log \left(\frac{P(x_i)}{1/n} \right) \\ &= \log n + \sum_{i=1}^{n} P(x_i) \log P(x_i). \end{aligned} \tag{4.14}$$

When $H(P,U)$ is to be minimized over all elements $P \in \mathcal{P}_K$, the first addend in (4.14) is a constant and can be neglected. Hence, we are looking for the probability distribution $P \in \mathcal{P}_K$ that has *maximal entropy*

$$H(P) = -\sum_{i=1}^{n} P(x_i) \log P_(x_i). \tag{4.15}$$

This procedure can be summarized thus:

> **Maximum Entropy Principle (MaxEnt):** If $\mathcal{P}_K$ is the set of probability distributions compatible with the agent's background knowledge K the agent should select the element $P \in \mathcal{P}_K$ that maximizes the entropy $H(P)$ as defined in (4.15).

It is easy to show that in the absence of specific background information, the uniform distribution maximizes entropy. In such a case, we are urged to choose the 'flat' distribution, i.e. the distribution that assigns equal probability $1/n$ to any elementary sentence. Similarly, when substantial information is available, we do not choose the uniform distribution itself, but the distribution that equivocates the probabilities as far as possible. In agreement with the coding-theoretic motivation, we choose the most uninformative

[29] Indeed, it would not make sense to average the loss according to the approximating density instead of the true density.

[30] Among the symmetric alternatives, there are the Hellinger distance $H(f,g) := \sum_{i=1}^{n}(\sqrt{f(x_i)} - \sqrt{g(x_i)})^2$ and the usual L^p-metrics $[\sum_{i=1}^{n}(f(x_i) - g(x_i))^p]^{1/p}$.

and least committing distribution among all admissible distributions. Thus, conditionalization is replaced by another form of belief revision. Objective Bayesians do not update their probability distributions by conditionalization – instead they choose the most equivocating probability distribution that is compatible with the available background knowledge. This strategy is called *cross-entropy updating* (because the cross-entropy to the uniform distribution is minimized). The advantage of cross-entropy updating consists in the fact that some information that cannot be processed in Bayesian Conditionalization is able to enter the game, e.g. information about the moments of the target distribution (that it has expectation μ, variance σ, etc.). The Principal Principle did not explain how this kind of information about objective chances affects rational credences. This is a serious drawback for the personalist position and underlines the flexibility of the objective Bayesian's cross-entropy updating.

The principle to choose probability values that are as middling as possible sounds attractive, but it requires substantiation. The standard idea is that middling, equivocating probabilities are per se less biased and less committal than extreme values (i.e. values near 0 and 1). But why should we avoid bias and commitment? Isn't a tendency towards middling probabilities not a form of bias, too? A possible justification consists in the idea that in the absence of further information, there is no justification for having a higher credence in a specific element of the sample space than in another one. All elements should be treated equally and be assigned the same probability unless background knowledge forces us to do so. This is the *Principle of Indifference* (Laplace [1814] 1951, Keynes 1921), a special case of the Maximum Entropy Principle. However, this kind of reasoning runs into objections well-known as varieties of *Bertrand's paradox*.[31] Assume that a factory produces miniature dice with a side length between zero and one centimeter. What is the probability (i.e. the rational credence) that a randomly selected die has a side length of more than half a centimeter? The Principle of Indifference suggests to answer '1/2' (and so does the Maximum Entropy Principle). So far, so good. Now, what is the probability that a randomly selected die has a *volume* of more than 0.125 cm^3? It is tempting to reply '1/2' again – the volume of the dice ranges between 0 and 1 cubic centimeter and we apply the Principle of Indifference. But our previous reply has already done the job – if the

[31] In my presentation of the paradox, I follow Hájek (2007).

probability of selecting a die with a side length longer than half a centimeter is equal to 1/2 then the probability of selecting a die with a volume greater than 0.125 cm^3 is also 1/2 and not 1/8. Two different applications of the Principle of Indifference yield two different probabilities for one and the same outcome. Everything depends on the quantity to which the Principle of Indifference is applied, and it is not clear that we can always give a 'natural' quantity of reference. Hence, neither the Principle of Indifference nor the Maximum Entropy Principle can be universally applied.

The Maximum Entropy Principle can be rescued from Bertrand's paradox by noting that there is a basic outcome space Ω. Entropy maximization merely applies with respect to the distributions of the members of Ω – and not to all possible reparametrizations of Ω as the preceding example illustrated. In the above counterexample, choosing such a Ω would also fix the parameter of interest (side length vs. volume). But then, the choice of the appropriate outcome space becomes a non-trivial and important task. To use the words of the above example: Shall we take side length or volume as the basic parameter relative to which we equivocate our degrees of belief? And if we have made a decision, how do we defend it? Making such decisions sounds like a highly arbitrary and subjective task. The problem echoes Nelson Goodman's (1983) point about formal theories of confirmation: whenever a hypothesis is confirmed according to 'reasonable' formal accounts of confirmation, several gerrymandered hypotheses which are interdefinable with the original hypothesis are equally confirmed. This was the upshot of the 'grue' paradox which we encountered in the first chapter. In the actual problem, all parameters (side length and volume) can be interchangeably used and isomorphically mapped onto each other so that it is not clear to which of them entropy maximization should be applied.

We have tried to escape Goodman's problem by delimiting the scope of formal theories of confirmation – formal accounts of confirmation *presuppose* a set of elementary, projectible predicates. But this does not entail that constructing such accounts is a futile activity. Confirmation theory merely has to be supplemented by a proper choice of elementary predicates. In a similar vein, we should adopt a charitable reading of entropy maximization: the principle does not hold unrestrictedly, but only relative to a reasonable parameter. The syntax of MaxEnt alone cannot determine which parameter is best for a problem at hand, in the same way that no formal account of confirmation can determine which predicates are projectible. Therefore it is somehow unfair to charge the objective Bayesian with Bertrand's paradoxes. Instead, it should be acknowledged that Maximum Entropy gives prescriptions for rational credences only relative to a suitable choice of the outcome space – the set of elementary propositions. The defense of that partition, however, depends on the specific case and is independent of the virtues and vices of the Maximum Entropy Principle.

This objection to objective Bayesianism being rebutted, I would like to scrutinize three arguments for preferring objective Bayesianism over a personalist position. The first argument stresses the objectivity and the lack of subjective dissent of objective Bayesianism: it is a commonplace in science to ask for objective inference and to have nothing but contempt and disdain for 'subjective' (i.e. personalist) theories of rational belief (see Dennis 2004 for an example). For example, after conducting a scientific experiment we are faced with the question which hypothesis we should rationally believe. For the personalist, this question cannot be conclusively answered since by adopting a sufficiently extreme prior opinion, any posterior credence may be rendered rational. This looks somehow undesirable since all conclusions and lessons from experience seem to be relative to prior opinion. For instance, policy-makers could reject inconvenient recommendations based on scientific insights by adjusting their prior beliefs – a strategy that is common in the global warming controversy, too. But the call for absolute objectivity has its drawbacks, too. Not all scientific applications require fully objective methods, and sometimes, it can be sensible to leave some wiggle room for subjective expert opinions. Such subjective expertise becomes the more important the more noisy and scarce our data. Furthermore, the objectivity argument would only show the need for *some* form of objective Bayesianism, but would not prove the superiority of entropy maximization over other objective Bayesian principles.

The second argument brings up the issue of efficiency. The elicitation of subjective beliefs is a time-comsuming activity that binds a lot of valuable human resources. People have to be asked about their credences, and if they shrug, it will be necessary to explain to them that they should imagine a fair bet, and so on. This is especially salient for AI applications where efficient resource allocation is especially important. For this reason, objective Bayesianism should be adopted, or so its proponents argue, so much the more as Bayesian nets nowadays greatly reduce the computational complexity of

Action/State	S	$\neg S$
A	3	-3
$\neg A$	0	-1

Table 4.1: A utility matrix in a decision problem.

entropy maximization.[32] – Similar to the first argument, this point stresses the drawbacks of a personalist position, but it does not make a case for a specific form of objective Bayesianism. Such a case is allegedly delivered by the third argument: the Maximum Entropy Principle is more *cautious* and *risk-averse* than other forms of objective Bayesianism (e.g. a principle that prescribes to *minimize* entropy). Of course, MaxEnt is not always more cautious and risk-averse than other constraints on rational credences, but due to its equivocating strategy, it rarely triggers a decision for which a high degree of certainty is required. Such decision are paradigmatic cases of 'risky' decisions (we will only make them if we feel very confident in the underlying facts). Thus, MaxEnt is *on average* more risk-averse than other constraints.[33]

Nevertheless I would like to bring up some arguments against the Maximum Entropy Principle, too. First, agents which are not risk-averse, but risk-neutral have no reason to prefer Maximum Entropy over other forms of objective Bayesianism. For such agents, the principle to minimize the entropy is on average as cautious as the Maximum Entropy Principle.[34] Indeed, such risk-neutral agents might exist in real life, too, for example in the investment department of high street banks. Second, there is a tension between the Max-

[32] See Williamson 2005.

[33] It might be argued that this result is not correctly stated. Here, a risky decision is understood as a decision that is executed if and only if the probability of a specific proposition S exceeds a very high value (e.g. take action A if and only if $P(S) > 0.9$). But it would be equally natural to define risky decisions as decisions where we expose ourselves to high potential losses. If we are expected utility maximizers, these definitions need not coincide – high potential losses can be outweighed by high potential gains, leaving the probability threshold for taking a decision unaffected. This is illustrated in table 4.1 – if the probability of S exceeds 1/4, the expected utility of action A is higher than the expected utility of action $\neg A$, although A is more risky in the sense that high potential losses threaten. If risk aversion in the sense of the Minimax Principle enters the considerations, i.e. if we want to avoid high potential losses under all circumstances, we will be less ready to take action A although objective Bayesians have no qualms with A. The argument for the Maximum Entropy Principle as opposed to other forms of objective Bayesianism works only if the first understanding of a risky decision ('high probability') is adopted.

[34] See Wiliamson 2007.

imum Entropy Principle and Bayesian Conditionalization. Recall that the Maximum Entropy Principle builds on cross-entropy updating: we choose the probability distribution that minimizes the cross-entropy to the uniform distribution over all distributions compatible with the available information. Seidenfeld (1979b, 1986) has argued that cross-entropy updating "commits the agent to more information than he/she may be entitled to"[35]. The problem occurs when nuisance parameters enter the field – contrary to Bayesian Conditionalization, cross-entropy updating neglects uncertainty about nuisance parameters and treats the distribution for the parameter of interest as a nuisance-free distribution. Instead of properly representing the uncertainty about the nuisance parameter, an 'ignorance' distribution is adopted for the parameter – but that is a crucial difference.[36] Third and last, allowing for subjective prior probabilities seems to threaten the objectivity of scientific inference, but in fact, it can also improve an inference. For instance, prior probabilities typically depend on subjective expertise of a scientist and not only on hard data. In an objective Bayesian framework, it would be impossible to bring to bear this expertise on the posterior probabilities. But if we adopt a personalist framework, several experts in a field are allowed to specify their prior beliefs and we can merge their opinions to obtain a conclusion that is based on the 'average' opinion of a group of experts. This merged posterior distribution may be more deliberate and precise than the objective Bayesian's posterior distributions. Objective Bayesianism does not make use of the scientists' knowledge about their field unless it is quantified in hard data. To be sure, it depends on the specific application whether an objective or a personalist approach should be preferred. Objective approaches can be implemented mechanically and are more efficient than subjective ones which require the elicitation of prior probabilities; furthermore they do not allow for subjective bias. But in a situation where the danger of manipulation and deliberate bias of prior probabilities is low, the personalist method incorporates a lot of subjective expertise which would otherwise get lost. Hence, the debate between objective Bayesian and personalist positions cannot be neatly decided the one or the other way. Both positions have their merits, and often, the nature of the specific inference problem decides the question.

[35] Seidenfeld 1979b, 433.

[36] Seidenfeld's (1979b) instructive example will be rehearsed in chapter six of the book. See, however, Williams 1980 and Uffink 1996 for attempts to reconcile MaxEnt with Bayesian conditionalization.

This survey of Bayesian positions is, of course, far from complete – in particular, we did not mention at all the concept of *logical probability* and *inductive probability* – a concept introduced by Rudolf Carnap (1950) as 'degree of confirmation' and subsequently developed by several logicians, among them Roberto Festa (1986, 1993) and Patrick Maher (2006). In Maher's reading, inductive probabilities do not correspond to degrees of belief – instead, inductive probability characterizes the relation between evidence and a hypothesis. It is the explicatum for the ordinary concept of probability in so far as this concept diverges from objective chance. Take the sentence

> "WB. The probability that a ball is white, given that it is either black or white, is 1/2."[37]

Maher continues:

> "Practically all competent speakers of ordinary language will assent to WB. The reference in WB to evidence, and the lack of any suggestion of a repeatable experiment, makes it clear that this is a statement of inductive probability, not physical probability (=objective chance, J.S.). Consequently, the truth value of WB does not depend on empirical facts but is determined by the relevant concepts."[38]

The idea is that statements as WB cannot be statements about objective chance. Ordinary competent speakers are inclined to consider WB to be true, although it is not about degrees of belief – it is about the relation between a hypothesis (the ball is white) and evidence (the ball is either white or black). Thus, we should introduce a concept of probability that is independent of both objective chance and subjective degrees of belief and where the truth value of probability statements merely depends on the meaning of its components (as in WB). In that sense, the concept of inductive probability is a logical one. Here, we see the difference to objective Bayesianism: objective Bayesianism constrains rational degrees of belief whereas inductive probability explicates a use of probability that is common in ordinary language. Of course, there are some striking similarities – in the above example, the objective Bayesian recommendation agrees with the inductive probability that the ball is white (=1/2). Therefore, both positions are often thought to

[37] Maher 2006, 195.

[38] Maher 2006, 195.

be closely related. But in fact, the projects of inductive probability and objective Bayesianism are quite different (although the probability *values* may agree): inductive probability is concerned with explication of a concept of ordinary language whereas objective Bayesianism cares for rationality constraints on degrees of belief. The motivation for objective Bayesianism is completely *a priori* and independent of any use of the word 'probability': several constraints on rational credences are introduced and motivated, and the MaxEnt summarizes them in a single principle. On the other hand, proponents of inductive probability are tied to the use of probability by ordinary speakers of a language. For instance, if nearly all speakers agreed on

> "WB. The probability that a ball is white, given that it is either black or white, is 1/3."

the concept of inductive probability would have to be modified. Unlike objective Bayesianism, it is a descriptive theory of the probability concept, not a normative theory of degrees of belief. For this reason, I refuse to classify the logical/inductive probability concept as a form of Bayesianism.

The above survey of positions gives a rough idea of the varieties of Bayesianism that are found in practice. For several aspects of Bayesian confirmation theory, the differences between the positions are negligible, e.g. the problem of finding a suitable measure of confirmation is equally pressing in all varieties of Bayesianism.

4.5 Summary

This chapter has introduced how probabilities express degrees of belief and how the calculus of probability is able to provide a logic of partial belief, in the same way that deductive logic offers a logic of full belief. Degrees of belief are explicated as hypothetical fair betting odds.

We have made the acquaintance of three principles – the Dutch Book theorem, the Principal Principle and Bayesian Conditionalization. The Dutch Book theorem shows why probabilities are the adequate mathematical tool for representing degrees of belief and that a set of bets that fail to conform to the axioms of probability cannot be fair. The Principal Principle calibrates rational credences with information about objective chances in real-world processes. Finally, Bayesian conditionalization determines how degrees of

Feature	Orthodoxy	Personalism	Maximum Entropy
intuitive	no	yes	medium
empirical constraints	no	yes	yes
efficient procedure available?	no	no	yes (Bayes nets)
subjective bias ruled out?	no	no	yes
merger of opinion possible?	yes	yes	no
especially risk-averse	no	no	yes
compatible with (BC)	yes	yes	no

Table 4.2: An overview over the advantages and drawbacks of the three main Bayesian conceptions of rational belief.

belief have to be changed in the light of incoming information that is regarded as certain. While the Dutch Book theorem and the Principal Principle impose constraints on synchronically held degrees of belief, Bayesian Conditionalization is a diachronic constraint.

There are a variety of Bayesian positions that agree only in the fact that the axioms of probability are unanimously accepted. Bayesian orthodoxy is the most parsimonious position, adding merely the principle of Bayesian conditionalization. Personalists moreover add the Principal Principle whereas objective Bayesians replace conditionalization by the Maximum Entropy Principle. Table 4.2 gives a survey over the properties of various Bayesian positions and illustrates their virtues and vices. Now we can proceed to the centerpiece of the book – the discussion of Bayesian confirmation theory.

Chapter 5

Bayesian Confirmation

In the last decades, probabilistic models and probabilistic methods have gained much ground in the empirical sciences, in particular the social sciences. Due to the increasing application of probabilistic modeling in science, confirmation theory should explicate the valid principles of probabilistic inductive reasoning, too. By definition, qualitative models of confirmation cannot accomplish that task. Moreover, they have struggled with some problems (e.g. the Duhem-Quine problem) that ask for a quantitative treatment of confirmation and for measures of *degree of support*. The last chapter has done some preliminary work in order to set up a probabilistic, quantitative theory of confirmation: we have become acquainted with a probabilistic calculus for degrees of belief. We have argued for specific constraints on a system of rational degrees of belief, in particular for immunity to Dutch Books and adherence to the Principal Principle. Furthermore we have introduced principles of belief updating as Bayesian conditionalization and entropy maximization, the latter being characteristic of an objective Bayesian framework. On these grounds, the Bayesian account equates confirmation with increase in rational degree of belief. However, such an explication does not automatically answer the central question – to which degree is a hypothesis confirmed by a piece of evidence? In this chapter, we discuss various measures of support and present some successes of Bayesian confirmation theory, as well as open questions.

5.1 Confirmation measures

Qualitatively, Bayesian confirmation amounts to an increase in rational degree of belief upon learning the new evidence E – in other words, evidence E confirms H if and only if $P(H|E) > P(H)$. But we have to remember a lesson from the very first chapter of the book – confirmation is a three place predicate, relative to background knowledge. As the Duhem-Quine problem teaches us, background assumptions are a crucial part of scientific reasoning and must be part of a proper confirmation theory. The natural way to integrate them consists in taking those assumptions for granted and in conditionalizing an agent's degrees of belief on them.[1] That said, we can write down a first, qualitative definition of Bayesian confirmation:

Definition 5.1 *A piece of evidence E* confirms *a hypothesis H relative to background assumptions K if and only if $P(H|E.K) > P(H|K)$.*

This definition gives a probabilistic explication of *relative confirmation*, not of *absolute confirmation*. Definition 5.1 describes the relevance of evidence for a hypothesis, not high credibility of a hypothesis. But it is not difficult to imagine a probabilistic condition for absolute confirmation, namely the condition that H enjoys a sufficiently high probability/rational credence. For instance, we could demand $P(H) > 0.99$ – then H would be (absolutely) confirmed 'beyond reasonable doubt'. Nevertheless, our main interest is devoted to the relationship between theory and evidence so that we focus on the incremental, relative concept of confirmation (=evidential relevance), as we did in our discussion of qualitative confirmation theory. Definition 5.1 is purely qualitative and leaves open which degree of support a piece of evidence E lends to a hypothesis H, relative to background K. Such a degree of support is required in order to tackle resilient challenges as the Duhem-Quine problem (see chapter 3): Duhem has rightfully argued that the test of a hypothesis is only as reliable and powerful as the auxiliary hypothesis that enter the testing process. But we would like to avoid the conclusion of *confirmational holism* that it is not meaningful to speak about the (dis)confirmation of single hypotheses. We require measures of support in order to show that in the case of experimental failure, the blame can be unevenly distributed over hypotheses under test and hypotheses in use.

[1] Nonetheless, for reasons of convenience, we will often speak (but not write) as if the background knowledge were empty.

Assume that a hypothesis H together with auxiliary hypothesis A predicts a certain phenomenon E, but the observations fail to produce E, falsifying the hypothesis. Then, on a Bayesian account, it is possible that the blame E puts on $H.A$ is not evenly distributed between H and A, i.e. for an adequate measure of support, the hypothesis is much more disconfirmed than the auxiliary hypothesis, or vice versa.[2] Recall Young's double slit experiment from chapter 2 that was set up to check the hypothesis that light propagates in waves (H). Among the auxiliary hypotheses is the assumption A that the lens used to focus the light behind the double slit is properly cut (see figure 2.1, p. 33). When conducting the experiment, we do not observe the characteristic inference pattern shown in figure 2.1, but instead a diffuse image (E). Clearly, this violates the predictions of H (light propagates wavelike), so that $P(E|H.A.K) = 0$. By contrast, if the lens is not properly cut, we will quite probably observe a diffuse image ($P(E|H.\neg A.K) = 4/5$). Moreover, we know that the quality of the available lenses varies a lot – only fifty percent of all lenses are properly cut. Hence $P(A|K) = 1/2$. Now we make the further (ad hoc) assumptions that $P(H|K) = 0.8$ (H is quite likely) and $P(E|\neg H.K) = 3/5$ (under the alternative hypotheses, it is not clear what we will observe). Some innocent independence assumptions with respect to A, H and K then yield

$$\begin{aligned} P(E|H.K) &= P(A|H.K)P(E|H.A.K) + P(\neg A|H.K)P(E|H.\neg A.K) \\ &= 1/2 * 0 + 1/2 * 4/5 \;=\; 2/5, \end{aligned}$$

and consequently,

$$\begin{aligned} P(E|K) &= P(H|K)P(E|H.K) + P(\neg H|K)P(E|\neg H.K) \\ &= 4/5 * 2/5 + 1/5 * 3/5 \;=\; 11/25. \end{aligned}$$

In a similar vein, we calculate that $P(E|A.K) = 4/25$. Combining that all with the help of Bayes's theorem yields

$$P(H|E.K) \approx 0.727 \qquad\qquad P(A|E.K) \approx 0.181$$

which indicates that A suffers much more under the failure of the experiment than H – recall that the prior probability of A was 1/2 and the prior

[2] Howson and Urbach (1993, 96-102) give a detailed and instructive example from the history of chemistry, but I think that the basic point can be illuminated more easily. Further discussion of Bayesian approaches to the Duhem-Quine problem takes place in Strevens 2001, 2005 and Fitelson and Waterman 2005, 2007.

probability of H was 4/5. This shows how the Duhem-Quine problem can be tackled in a Bayesian framework: both hypothesis are disconfirmed, but not to an equal degree. However, while we intuitively grasp that H is less undermined than A, we require a measure of support to make this reasoning precise. Therefore, this chapter revolves around the comparison of various measures of support. In order to unify and to normalize the various measures, we demand that all measures of confirmation $\mathfrak{c}(H, E, K)$ satisfy

$$\mathfrak{c}(H, E, K) \begin{cases} > 0 & E \text{ confirms } H \text{ relative to } K \\ = 0 & E \text{ is neutral to } H \text{ relative to } K \\ < 0 & E \text{ disconfirms } H \text{ relative to } K. \end{cases} \tag{5.1}$$

A measure of support takes a positive value if and only the evidence confirms the hypothesis and a negative value if and only if the evidence disconfirms the hypothesis.

There is an abundant list of confirmation measures proposed in the literature, and discussing them all would be too tedious even in a monograph on confirmation and evidence. Instead, I follow Fitelson's (2001a, 2001b) strategy to select some measures of support that are (1) present in the confirmation-theoretic debate (2) defended by a non-negligible part of the research community (3) representative of different approaches to quantify confirmation. Luckily, many measures can be eliminated from the list because they are *ordinally equivalent* to one of the measures in the list. Ordinal equivalence amounts to the following condition:

Definition 5.2 *Two confirmation measures $\mathfrak{c}_1$ and $\mathfrak{c}_2$ are ordinally equivalent if and only if for all H, H', E, E', K and K':*

$$\mathfrak{c}_1(H, E, K) \geq \mathfrak{c}_1(H', E', K') \Leftrightarrow \mathfrak{c}_2(H, E, K) \geq \mathfrak{c}_2(H', E', K'). \tag{5.2}$$

In other words, two ordinally equivalent measures may assign different degrees of support to a tuple $\langle H|E|K \rangle$ as long as they impose the same partial order. Thus, they agree with regard to the ordinal structure of inductive support, i.e. with regard to judgments which hypotheses are better or equally confirmed by pieces of evidence. In other words, by 'stretching and contracting' the measures it is possible to map them onto each other. The different degrees of support merely remain a matter of appropriate scaling so that ordinally equivalent measures share most interesting properties.

For those reasons, we will only pick a single representative of a particular class of ordinally equivalent measures.[3] The six classes of measures which are now introduced represent the most popular and fiercely discussed suggestions in confirmation theory and they also cover a significant part of the ordinal structure that a measure of confirmation could possibly impose.[4]

Difference Measure

$$d(H,E,K) := P(H|E.K) - P(H|K)$$

Log-ratio Measure

$$r(H,E,K) := \log \frac{P(H|E.K)}{P(H|K)}$$

Counterfactual Difference Measure

$$\begin{aligned} s(H,E,K) &:= P(H|E.K) - P(H|\neg E.K) \\ &= \frac{P(H|E.K) - P(H|K)}{P(\neg E|K)} \end{aligned}$$

Likelihood Ratio Measures

$$l(H,E,K) := \log \frac{P(E|H.K)}{P(E|\neg H.K)} \qquad \text{(Log-Likelihood)}$$

$$k(H,E,K) := \frac{P(E|H.K) - P(E|\neg H.K)}{P(E|H.K) + P(E|\neg H.K)} \qquad \text{(Kemeny-Oppenheim)}$$

Covariance Measure

$$\begin{aligned} c(H,E,K) &:= P(H.E|K) - P(H|K)P(E|K) \\ &= P(H|K)[P(E|H.K) - P(E|K)] \end{aligned}$$

Crupi's and Tentori's z-measure

$$z(H,E,K) := \begin{cases} \frac{P(H|E.K)-P(H|K)}{1-P(H|K)} & P(H|E.K) \geq P(H|K) \\ \frac{P(H|E.K)-P(H|K)}{P(H|K)} & \text{otherwise} \end{cases}$$

[3] For historical reasons, a single exception (the log-likelihood-measure and the Kemeny-Oppenheim measure) will be made.

[4] Further suggestions include the measures of Gaifman (1979), Nozick (1981), Mortimer (1988) and Rips (2001).

Some explanations are due, of course. The difference measure, advocated by Earman (1992) and Rosenkrantz (1994), is an intuitive measure that takes the difference between the posterior and the prior degree of belief in H as a measure of the support evidence E lends to H. – The log-ratio measure r, proposed by Howson and Urbach (1993) and Milne (1996), replaces the difference between the posterior and the prior degrees of belief in H by the (logarithmic) ratio of those two quantities. So r can potentially take very high and low values whereas d is restricted to the interval $[-1; 1]$. – The counterfactual difference measure s is a variation of the difference measure that compares the posterior degree in H with the degree of belief that we would put in H *had* $\neg E$ *occurred instead of* E. In a simplified manner of speaking, we could also say that s multiplies the difference between prior and posterior credence in H with the expectedness of the evidence. That measure is most prominently backed by Christensen (1999) and Joyce (1999). – The log-likelihood measure l takes the logarithmic ratio of the likelihoods of the evidence, once under H and the other time under $\neg H$, as a measure of support. It is ordinally equivalent to the Kemeny-Oppenheim measure k so that the latter does not need special attention. However, it is noteworthy, first for historical reasons, and second, because it is normalized to [-1; 1]. Adherents include Kemeny and Oppenheim (1952), Good (1983), and Fitelson (2001a, 2001b). More on this later – c is a variation of Carnap's (1950, §67) relevance measure that takes the degree to which H and E are correlated (in other words, the *covariance* of E and H) as a measure of support. – Finally, z is a recent proposal by the philosopher Vincenzo Crupi and the psychologist Katya Tentori that is backed by ordinary people's judgments about degree of support (Tentori et al. 2007). Furthermore it has some theoretical symmetry properties which single it out among all confirmation measures (Crupi et al. 2007; Crupi et al. 2008).

Of course, all measures satisfy minimal adequacy constraints in the sense that they satisfy equation (5.1) (proofs omitted). Moreover, several measures found in the literature differ from the above list only by means of adding or omitting a logarithm symbol – e.g. there is a non-logarithmic version of the log-ratio-measure r. Since the logarithm is a monotone function, such measures are ordinally equivalent to measures in the list so that we can neglect them.

If the measures are compared, the first criterion that could be used is a kind of intuitive plausibility. But we should not rely too much on intuitions when we compare two measures in terms of intuitive adequacy. For instance, both the difference measure d and the log-ratio measure r are based on the increase in probability from $P(H|K)$ to $P(H|E.K)$. The log-ratio measure seems to correctly capture that the increase in H's probability from 0.000001 to 0.01 is far more significant than the increase in probability from 0.24 to 0.25. The difference measure yields the opposite result which we find unintuitive. But on the other hand, the log-ratio measure also asserts that an increase in probability from 0.001 to 0.01 lends more support to H than an increase in probability from 0.1 to 0.9.[5] Here, the difference measure seems to be in agreement with our intuitions. This dilemma seems to show that both measures cannot be adequate explications of the concept of inductive support – and indeed, we will later encounter arguments for this claim. But we should not be too hasty. For each intuitive example which seems to favor a specific measure over another one, there might be other examples which argue for the opposite claim. It will not be possible to account for all intuitions that are connected with confirmation, evidential relevance and inductive support. Rather, an adequate explication of these vague concepts will also be assessed in terms of fruitfulness and precision whereas some intuitions will be preserved and others will be abandoned. Christensen (1999, 438-39) presents a nice analogy that illustrates the problems of finding an adequate measure of inductive support: how do we measure the extent to which a politician P (e.g. someone who runs for the American presidency) is financially supported by a group G? The proportion of G-donations in P's funds? P's relative position in the presidential run as a function of the G-donations? And so on. Christensen conjectures

> "Thinking about these different measures of support suggests to me that there is no single clearcut question being asked when we ask 'How much support does P get from G?'. It would not be surprising if the same were true of the question 'How much does evidence E support hypothesis H [relative to K, J.S.]?'.[6]

Nevertheless, I do not endorse Christensen's pessimistic suggestion. Admittedly, it is unlikely that a measure of confirmation will ever succeed to capture all intuitions that are connected to the concept of confirmation. But nevertheless, it is possible and worthwhile to narrow down the list of proposed

[5] See Christensen 1999, 438.

[6] Christensen 1999, 439. Notation changed for convenience.

measures by a list of clear-cut criteria: first, their ability to resolve classical problems of confirmation theory, as the problem of irrelevant conjunctions. Second, the satisfaction of appealing theoretical properties and the absence of undesirable properties. Third, the explanatory power with respect to features of scientific practice, as the power of independent or surprising evidence. Indeed, we will see that many problems in confirmation theory are *measure sensitive*, i.e. whether a problem remains or vanishes depends on the specific measure which is used. The criteria which I set up are quite uncontentious and liberal, so that a measure of confirmation that does not satisfy them is really in trouble, independent of the specific application context. In the end three measures will remain possible candidates for an explication of inductive support. Finally, these three measures are discussed and compared to each other.

5.2 Adequacy conditions on measures of support

Finding a suitable set of adequacy criteria for a measure of support is no easy task, so much the more as the proposed measures exemplify various ways to characterize the concept of confirmation. Indeed, there are two grand traditions in quantitative confirmation theory. On the one hand, we have measures like d, r and s which try to capture the increase in degree of belief. On the other hand, measures like k, l and z quantify the strength of an inductive argument which the evidence gives in favor of the hypothesis (relative to the background knowledge). These two traditions fundamentally disagree on certain aspects of inductive support, as we will soon see. So it is worthwhile to see which measure (and which tradition) is more capable to solve the various open problems.

One of the very basic desiderata on a measure of support consists in accounting for the power of surprising evidence. When various pieces of evidence are equally likely under a certain hypothesis, it seems that an *unexpected* piece of evidence E confirms the hypothesis to a stronger degree than an expected piece of evidence E'. The hypothesis takes more risk in predicting E than it takes in predicting E'. Theory testing is a *contrastive* activity and E brings out the contrast between H and $\neg H$ to a stronger degree than E'. Therefore, it should confirm it to a higher degree, too, as witnessed by the following equations:

$$P(E|K) = P(E|H.K)P(H|K) + P(E|\neg H.K)P(\neg H|K)$$
$$P(E'|K) = P(E'|H.K)P(H|K) + P(E'|\neg H.K)P(\neg H|K).$$

If the first addend on the right hand side is the same for both equations (which is guaranteed by assumption), the condition $P(E|K) < P(E'|K)$ implies that E must be less likely under $\neg H$ than E' – $P(E|\neg H.K) < P(E'|\neg H.K)$. Thus E has a higher contrastive power than E'. Indeed, in famous cases of confirmation in science the observation of an unexpected event often plays a crucial role: Eddington's unexpected verification of Einstein's predictions about the bending of starlight by the sun was salient evidence for the General Theory of Relativity as opposed to Newtonian mechanics. Indeed, the virtue of surprising evidence is vindicated by all measures of support on our list:

Fact 5.1 *Assume that $P(E|H.K) = P(E'|H.K)$, $P(E|K) < P(E'|K)$, and that both E and E' confirm H relative to K. For all presented measures of confirmation $\mathfrak{c} \in \{d,r,l,k,s,c,z\}$, E confirms H to a higher degree than E' (relative to K).*

Proof: trivial for all measures but s. For s, note that by Bayes's theorem,

$$s(H, E, K) = P(H) \left[\frac{P(E|H.K)}{P(E|K)} - \frac{1 - P(E|H.K)}{1 - P(E|K)} \right]. \tag{5.3}$$

For fixed $P(E|H.K) = P(E'|H.K)$, the term in the brackets in (5.3) is monotonously decreasing in $P(E|K)$. (Calculating the first derivative easily proves that.) Therefore $s(H, E, K) > s(H, E', K)$. □

The next problem – the problem of irrelevant conjunctions – will be more measure-sensitive. Hypothetico-deductive confirmation has struggled with irrelevant conjunctions – if a hypothesis H is confirmed, the conjunction of H and an arbitrary hypothesis X with which H is consistent is confirmed, too. Since X need not be relevant to the evidence – it can actually be an *arbitrary* hypothesis –, this result is unbearable. So we demand that a quantitative account of confirmation must solve or at least mitigate that problem.

In a Bayesian framework, we obtain that, if E confirms H relative to K, E does not necessarily confirm $H.X$ relative to K, for an arbitrary X. The conditions for Bayesian confirmation are $P(H|E.K) > P(H|K)$ respectively $P(H.X|E.K) > P(H.X|K)$, and neither inequality implies the other one. For instance, E can be positively (probabilistically) relevant to H, but negatively relevant to X, relative to K. Thus, the problem of irrelevant conjunction does not arise in its classical form.

Still, Bayesians are not out of trouble. Assume that H deductively entails E. In such a case $H.X$ will entail E for an arbitrary X. This special version of the problem of irrelevant conjunctions posed serious problems for a deductive theory of qualitative confirmation (see chapter 2). Deductive confirmation is a special case of Bayesian confirmation: From $H.K \models E$ it follows that $P(E|H.K) = 1$ which, by an application of Bayes's theorem, leads to

$$P(H|E.K) = \frac{P(H|K)P(E|H.K)}{P(E|K)} = \frac{P(H|K)}{P(E|K)} > P(H|K)$$

for any non-trivial evidence. So evidence that is logically entailed by the hypothesis always confirms it. Since logical entailment is preserved under strengthening the premises it follows that

> If $H.K \models E$ then E confirms $H.X$ in the Bayesian sense relative to K for any X that is consistent with $H.K$.

Hence, the problem of irrelevant conjunctions persists in a Bayesian framework, too – namely in the special case that the evidence is logically implied by the hypothesis. This result can be extended, as Hawthorne and Fitelson show in their 2004, namely to all conjuncts X that do not change the likelihood of E under H.[7] Hawthorne and Fitelson prove this for the measures d, l and r, but actually, their observation can be generalized to a result that holds for all Bayesian measures of confirmation that satisfy the minimal constraint (5.1):

Proposition 5.1 *Assume that E confirms H relative to K according to definition 5.1 and that $P(E|H.X.K) = P(E|H.K)$ for a sentence X consistent with $H.K$. Then E confirms $H.X$, too (and hence, all proposed confirmation measures give values greater than 0).*

Proof: The crucial condition $P(H|E.K) > P(H|K)$ is equivalent to $P(E|H.K) > P(E|K)$ as an application of Bayes's theorem makes clear. Therefore, due to $P(E|H.X.K) = P(E|H.K)$ we can infer that $P(E|H.X.K) > P(E|K)$ so that E confirms $H.X$ relative to K according to definition 5.1. Since all

[7] See also Fitelson 1999, 2002 for precursors.

measures of support on our list satisfy (5.1), they yield confirmation values greater than zero.□

This states the problem: the problem of irrelevant conjunctions persists if tacking an irrelevant conjunct to the hypothesis does not change the likelihood of the evidence. Indeed, if

$$P(E|H.X.K) = P(E|H.K) \tag{5.4}$$

this seems to be a typical case of a tacked hypothesis X which is *irrelevant* to E. For instance, the likelihood of tossing 'heads' with a coin is not changed by taking into account the additional information that all ravens are black. Such a kind of additional information has no impact at all on the outcome of the coin flip. Similarly for the converse – if additional information changes the likelihood of the evidence, it cannot have been completely irrelevant.[8] Cases in which $P(E|H.X.K) = P(E|H.K)$ are just the probabilistic version of tacking an irrelevant conjunct X. Therefore, this situation is the probabilistic counterpart of the classical problem of irrelevant conjunctions. Hence, the Bayesian approach does not give a straightforward *resolution* of the problem, contrary to what we might have hoped. True, the probabilistic concept of confirmation is wider than the concept of qualitative confirmation, and tiny probabilistic support already counts as confirming evidence. Therefore it is not so disturbing that irrelevant conjunctions get confirmed – probabilistic confirmation is more liberal than qualitative confirmation which focuses on structural relationships between evidence and hypothesis. This observation somehow mitigates the pull of the problem. Still, we would like to have a result showing that irrelevant conjunctions satisfying (5.4) are confirmed to a lower degree than the original hypothesis.

Hawthorne and Fitelson (2004) prove such a theorem for some measures of confirmation: the degree of confirmation is lowered by tacking irrelevant conjunctions that satisfy some minimal conditions. Specifically, they demonstrate that effect for the difference measure d and the log-likelihood measure l whereas the degree of support remains constant under the log-ratio-measure r. I amend this result by noting that the degree of confirmation is lowered by tacking irrelevant conjunctions if measured according to the other three measures suggested (s, c and z).

[8] See Fitelson 2002.

Proposition 5.2 *Assume that E confirms H relative to K and $P(E|H.X.K) = P(E|H.K)$ for a sentence X where $P(X|H.K) \neq 1$. Then the degree of confirmation E lends to H exceeds the degree of confirmation E lends to $H.X$, for the confirmation measures d, l, s c, z. For instance, $d(H,E,K) > d(H.X,E,K)$. But for the log-ratio measure r, it holds that $r(H,E,K) = r(H.X,E,K)$.*

Proof: For d, l and r see the proof of Theorem 2 in Fitelson and Hawthorne (2004). Furthermore, it is easy to see that

$$c(H,E,K) = P(H|K)[P(E|H.K)/P(E|K) - P(E|K)].$$

This entails that

$$\begin{aligned} c(H.X,E,K) &= P(H.X|K)[P(E|H.K)/P(E|K) - P(E|K)] \\ &\leq P(H.X|K)[P(E|H.K)/P(E|K) - P(E|K)] \\ &< c(H,E,K). \end{aligned}$$

The proof for s directly follows from the proof for d, due to $s(H,E,K) = d(H,E,K)/P(\neg E|K)$. For the proof for z, note that $z(H,E,K) = d(H,E,K)/[1-P(H|K)]$. Then, the proof for z again relies on the proof for d:

$$z(H.X,E,K) = \frac{d(H.X,E,K)}{1-P(H.X|K)} < \frac{d(H,E,K)}{1-P(H|K)} = z(H,E,K).$$

□

Hence, most confirmation measures mitigate the paradox of irrelevant conjunctions in so far as the degree of confirmation is lowered by tacking irrelevant hypotheses where the 'irrelevancy' of the tacked hypotheses is explicated by means of constancy in the likelihood (see equation (5.4)). These results thus provide an argument against the log-ratio measure r since the problem of irrelevant conjunctions clearly persists for this measure: tacking irrelevant conjunctions does not alter the degree of confirmation. According to r, the observation of a black raven confirms the hypothesis that all ravens are black and all doves are white to an equal degree as the original hypothesis that all ravens are black. That is unacceptable.[9] There are, however, also some arguments for r, and before dismissing r once and for all, we should

[9] This property of r, often called 'deductive insensitivity', was already discovered by Rosenkrantz (1981) and Gillies (1986).

listen to them. The most ambitious and famous argument is due to Peter Milne (1996). He sets up a list of desiderata for a measure of support and demonstrates that r is the unique measure of support which satisfies them all. Fortunately, we need not rehearse his argument in detail – instead we merely note that it relies on the requirement that

$$\text{if}\, P(E|H.K) = P(E|H'.K)\, \text{then}\, \mathfrak{c}(H, E, K) = \mathfrak{c}(H', E, K) \tag{5.5}$$

for a suitable measure of confirmation $\mathfrak{c}$. This constraint is, however, rather strong and unmotivated, as Fitelson notes in his 2001a. Fitelson makes a good case against Milne's argument, but I think (5.5) can be rejected even more directly, by recalling the problem of irrelevant conjunctions. In fact, if H' is H plus an irrelevant conjunct (i.e. $H' = H.X$ for an arbitrary X) we would very much want (5.5) *not* to hold. (5.5) just states the deductive insensitivity of a confirmation measure, and any such measure falls prey to the problem of irrelevant conjunctions: the degree of confirmation does not change when tacking an arbitrary X. But clearly, the degree of support should decrease when arbitrary many irrelevant conjuncts are tacked to the evidence. Hence, Milne's argument does not rise a new point in favor of r – one of its premises is unacceptable, as we have already seen. In particular, none of the other measures of support satisfy (5.5) and are thus immune to the problem of irrelevant conjunctions in its strongest form. Thus, r can be ruled out and five measures (more precisely, five ordinal classes of measures of support) stay in the game.

Next, we investigate the symmetry properties which the measures possess. Although symmetry properties are certainly attractive from a mathematical point of view, I do not want to postulate that a specific kind of symmetry has to be satisfied by an adequate confirmation measure. Of course, symmetries can be very valuable (as they often are in physics), contributing to the mathematical tractability and beauty of the underlying theory. But violating one of these 'beautiful' symmetries (e.g. in favor of 'approximate symmetries') does not constitute a knockdown argument against a measure of support. I believe it to be more fruitful to focus on vicious symmetries – symmetries which a measure of inductive support should not exhibit under any circumstances. Since it is much easier to argue on conceptual grounds for the *inadequacy* of a symmetry property, such symmetries serve as a means of detecting inappropriate measures of support. In their 2002, Eells and Fitelson argue that an adequate measure of support should *not* exhibit two particular symmetries, perspicuously demonstrated in the special cases $E \models H$ and $H \models E$. The first symmetry is *commutativity symmetry* and asserts that the degree of support E lends to H equals the degree of support H lends to E (both relative to K). This can be stated thus for a confirmation measure $\mathfrak{c}$:

$$\mathfrak{c}(H, E, K) = \mathfrak{c}(E, H, K). \tag{5.6}$$

Now, assume that a friend of yours is drawing cards from a standard 52-cards deck and the hypothesis H asserts that the uppermost card is a king of spades. Your friend draws that card and tells you that it is a black card. That seems to support the hypothesis that the uppermost card is a king of spades merely to a very moderate degree – there are many black cards in the game. Now consider the opposite situation: we hypothesize that the card is a black card and we observe that the card is a king of spades. This observation conclusively confirms the hypothesis that the card is black! The evidence in the latter case is much more informative than in the former case, giving decisive information for judging the correctness of H. Therefore commutativity symmetry is an undesirable property of a confirmation measure, so much the more as similar examples can be construed. Apart from that, inductive inference is often taken to generalize deductive inference and logical entailment, and clearly, logical entailment is no symmetric relation (i.e. $A \models B$ does not imply $B \models A$). Thus, conceptual reasoning and practical examples both suggest that confirmation is no symmetric notion. Hence, measures of support that satisfy (5.6) are problematic.

Second, I would like to turn to *evidence symmetry.* This symmetry asserts that if E confirms H to degree x (relative to K), $\neg E$ *disconfirms* H to the same degree:

$$\mathfrak{c}(H, E, K) = -\mathfrak{c}(H, \neg E, K). \tag{5.7}$$

Put verbally, the positive impact E exerts on H equals the negative impact $\neg E$ would have had (if observed). Negating the evidence changes the algebraic sign of the degree of support. – However, this property is not desirable either. Assume that the hypothesis that all ravens are black (H) is well-confirmed in the absolute sense, by a plethora of observations of ravens and non-ravens. Suppose that we observe a further object of which we only know that it is no counterexample to the raven hypothesis. Call this piece of evidence E. Since E rules out a further potential counterexample to H, this observation makes H more credible, if only to a minute degree. But what would have happened if $\neg E$ had been observed? We would have observed a non-black raven, conclusively falsifying H, in spite of all the support that previous observations have lent to H. It is clear that the amount of disconfirmation we get by observing $\neg E$ greatly exceeds the minute support we get by observing E.[10] The rationale behind that judgment is not difficult: evidence can be quite unspecific and uninformative (such as E in the above example), and observing the negation of unspecific evidence amounts to observing highly specific and informative evidence. Therefore no sensible measure of support exhibits evidence symmetry.

How do the various measures of support fare with respect to the symmetry constraints? The answers are easily calculated:

Proposition 5.3 *(Eells and Fitelson 2002)*

- *Among the confirmation measures d, r, k, l, s, c and z, only c and r satisfy commutativity symmetry* (5.6)*.*
- *Among the aforementioned measures, only c and s satisfy evidence symmetry* (5.7)*.*

Proof: trivial.

Hence, Eells's and Fitelson's findings give us a good reason to reject the measures c, s and r. Measures that satisfy (5.6) and/or (5.7) in all conceivable circumstances are certainly no good measures of support since the crucial asymmetries regarding the hypothesis/evidence relationship and the specificity of the evidence are neglected. Proposition 5.3 thus confirms that r is no adequate measure of support (see the problem of irrelevant conjunctions) and besides, it gives us a knockdown argument against c and s. It is also interesting to note that z does not only violate the problematic symmetries, but that it is the only measure that satisfies all (and only those) symmetries which are satisfied by deductive entailment (Theorem 2 in Crupi et al. 2007, 241). For instance, if E confirms H to degree x, $\neg H$ will confirm $\neg E$ to degree x, too, analogous to the law of contraposition. This makes z a natural candidate for an explication of inductive support in the tradition of inductive logic: confirmation and support are considered to be extensions and generalizations of deductive entailment. We keep this property in mind

[10] See section 3 in Eells and Fitelson 2002.

and meanwhile, we proceed to the next criterion: the Laws of Likelihood and their relation to the confirmation measures.

The Laws of Likelihood express sufficient conditions for the case that a hypothesis H is better confirmed than a hypothesis H' by the very same piece of evidence E. The strongest formulation of such a law can be found in Hacking (1965, chapter V); but for our purposes it will be more convenient to work with a paraphrase by Richard Royall (1997):

> **Strong Law of Likelihood (SLL):** If an event E is more likely under hypothesis H than under hypothesis H', then H is better supported by E than H' (relative to K). In other words, for a suitable measure of support $\mathfrak{c}$,
>
> $$P(E|H.K) > P(E|H'.K) \qquad \Rightarrow \mathfrak{c}(H,E,K) > \mathfrak{c}(H',E,K).$$

The idea of SLL can be expressed thus: the degree of support is a function of the degree to which the observed evidence is rendered likely by the various hypotheses at stake. The more the evidence is likely under a hypothesis, the better it supports the hypothesis. The main advantage of the SLL is the fact that these likelihoods are easy to compute and do usually not require assignment of prior probabilities. That paves the way for an objective theory of inductive inference. On the other hand, endorsing the Strong Law of Likelihood is not unproblematic either: First, assume that the confirming evidence E is slightly more probable under H than under H'. Assume further that H is a negligible hypothesis with a very low prior probability whereas H' is one of our best candidates for the true hypothesis. Then, the observed evidence will not significantly change our epistemic attitude towards H – we might be slightly more inclined to take it seriously, but still, it will remain a very improbable hypothesis. On the other hand, E could significantly favor H' over its main rivals that are not confirmed by E. The evidence seems to make a real epistemic difference for H', but not so much for H. This speaks against the SLL. Second, we may encounter the case $H' \models E \models H$. Here, H is logically entailed by the observed evidence, but the SLL asserts that H' is always more strongly supported than H. This contradicts an important intuition about confirmation: the inductive argument in favor of H has maximal strength, but still, the degree of support will always be lower than in the case of observing deductive predictions.[11] Third and last, note the

[11] See Fitelson 2007.

similarity to (5.5): If we make the plausible requirement that the measure of support be a continuous function, then it follows from the SLL that if $P(E|H.K) = P(E|H'.K)$, then $\mathfrak{c}(H, E, K) = \mathfrak{c}(H', E, K)$.[12] This was just the content of (5.5). In other words, the SLL makes us believe that the degree of support does not change by tacking irrelevant conjunctions, certainly an undesirable side effect. In total, we have listed three severe drawbacks of the SLL. Admittedly, one might object that the latter two arguments against the SLL are misguided: the proponents of the SLL wanted to apply it to the case of mutually exclusive hypotheses H and H'. This rules out the cases $H' \models E \models H$ and $H' = H.X$. But first, this form of the SLL is never explicitly articulated in the literature, second, it is a substantial weakening of SLL, and third, this modification does nothing to address the first criticism.

For these reasons, the SLL is usually considered to be too strong a requirement for measures of support. A less contentious requirement is the so-called Weak Law of Likelihood, which putatively captures an essential message of Bayesian confirmation theory:[13]

> **Weak Law of Likelihood (WLL):** If $P(E|H.K) > P(E|H'.K)$ *and* $P(E|\neg H.K) \leq P(E|\neg H'.K)$, then H is better supported by E than H', i.e. for a suitable measure of support $\mathfrak{c}$, $\mathfrak{c}(H, E, K) < \mathfrak{c}(H', E, K)$.

Why is the WLL so plausible? Joyce (2003) argues that the two assumptions of the WLL capture that H has uniformly higher predictive value than H'. On the one hand, the actual evidence E is better predicted by H than by H' because the likelihood of E is higher under H. On the other hand, if H had not been true, the actual evidence would have been less likely than if H' had not been true. Those two properties establish the predictive superiority of H. Put another way, E highlights the *contrast* between H and $\neg H$ to a stronger degree than the contrast between H' and $\neg H'$, generalizing the considerations with respect to surprising evidence made at the beginning of that section. This can be represented by the inequalities

$$P(E|H.K) \geq P(E|H'.K) > P(E|\neg H'.K) \geq P(E|\neg H.K) \tag{5.8}$$

[12] If E is more likely under H than under H', then the inequality $\mathfrak{c}(H, E, K) \geq \mathfrak{c}(H', E, K)$ holds. In the opposite case (E more likely under H' than under H) it is reversed, hence for the case $P(E|H.K) = P(E|H'.K)$ the degree of support must be equal.

[13] See Joyce 2003.

which are valid in case of confirmation. I would like to illuminate the plausibility of the WLL with an example that satisfies (5.8). We find out that Tom has an academic degree (E). This is better evidence for the claim that exactly one of his parents has an academic degree (H) than for the claim that none of his parents has an academic degree (H'): First, parents with exactly one academic degree are more likely to have children who will obtain academic degrees themselves than parents without any academic degrees, due to the impact of social background on educational career. Second, if *at least* one of Tom's parents had an academic degree ($= \neg H'$), it would be more likely that Tom has an academic degree than if either both or none of his parents had an academic degree ($= \neg H$). The reason is the same – academic background in the family positively affects the educational career of the children. For this reason, H is better supported than H'.

Clearly, the above example and the informal motivation underlying the WLL demonstrate that conformity to the WLL is no vicious property for a measure of support. But the opposite direction is contentious: Should measures of support be rejected because they fail to satisfy the WLL? I believe that this tenet goes too far. The above example examines a case where H and H' are mutually exclusive. In examples where H and H' overlap, it is much more challenging to appreciate the intuitive force of the WLL and to construct convincing examples. So we have to invoke the theoretical justification of WLL presented before the inequalities in (5.8). But even this is more difficult since H and H' share common content, making it harder to argue for the intuitiveness of the WLL. So it can rightfully be asked whether the theoretical argument in favor of WLL already *presupposes* a certain understanding of confirmation, namely in the sense of highlighting a contrast in the likelihoods (see again (5.8)). We have not yet presented any arguments for such an understanding which is, for instance, captured in the log-likelihood measure l. Certainly, there is some aesthetic and theoretical appeal in endorsing the WLL for the case of overlapping hypotheses, too, but it is much more difficult to *argue* for the WLL in this case because our intuitions are quite blurred. The WLL is often endorsed as a general rule because it can be so easily motivated for the case of mutually exclusive hypotheses, and in most motivations of the WLL (e.g. Joyce 2003), such a tacit assumption is made. For that case, the motivation is indeed clear-cut and plausible. Therefore I believe, contra Fitelson (2007) and Joyce (2003), that merely for mutually exclusive hypotheses, there is a compelling reason to accept the WLL as an adequacy criterion for measures of confirmation. Hence we replace WLL by

> **Weak Law of Likelihood (WLL), exclusive version:** If $P(E|H.K) > P(E|H'.K)$ *and* $P(E|\neg H.K) \leq P(E|\neg H'.K)$, and $\models \neg(H.H')$, then H is better supported by E than H', i.e. for a suitable measure of support $\mathfrak{c}$, $\mathfrak{c}(H,E,K) < \mathfrak{c}(H',E,K)$.

Luckily, for our purposes, the distinction between the two versions of WLL is not significant: All remaining confirmation measures satisfy WLL as well as the disjoint version of WLL.[14] Hence, for practical reasons, i.e. for adjudicating between the various measures of support it is not important whether we endorse only the disjoint or also the more general version of the WLL. Nearly all proposed measures satisfy the WLL. Only the log-ratio measure r satisfies the more demanding Strong Law of Likelihood whereas the other measures violate the SLL. This agrees with our former judgments regarding the inadequacy of r as a measure of support.

Another criterion of adequacy that has been proposed in the literature is the ability to deal with independent evidence. We often want to say that two pieces of evidence that are 'disconnected' from each other do jointly support a hypothesis more than each single piece does. This intuition is quite sound: if the support inherent in E_1 is not mediated via E_2 and does not depend on E_2 and vice versa, then the positive impact of $E_1.E_2$ on the rational credence in H exceeds the positive impact of a single piece of evidence. This line of reasoning can be defended in a counterfactual way, too: If the hypothesis were wrong, independent pieces of evidence would give us a better chance to spot the errors in the hypothesis than just one piece of evidence. This is a paradigm case for *independent evidence* or *evidential diversity.*

Nevertheless, it is a non-trivial task to give a satisfactory explication of evidential diversity. In his 2001b, Branden Fitelson compares different attempts to formalize this concept. Obviously, probabilistic independence is a good candidate for capturing evidential diversity (and independent evidence). Then, the question arises whether we should aim for conditional or unconditional independence. Like Fitelson, I opt for independence conditional on the hypothesis H, but my reasons are somewhat different: First, note that either H or $\neg H$ will be the case. Naturally, the relevant form of independence between two pieces of evidence is independence in both possible cases

[14] See Joyce 2003 and Fitelson 2007.

– H and $\neg H$. In probabilistic notation, this amounts to

$$\begin{aligned} P(E_1.E_2|H.K) &= P(E_1|H.K)\,P(E_2|H.K) \\ P(E_1.E_2|\neg H.K) &= P(E_1|\neg H.K)\,P(E_2|\neg H.K), \end{aligned} \tag{5.9}$$

and that is indeed the kind of independence which we mean when we speak about independent trials in testing statistical hypotheses. Assume, for instance, that we draw balls with from an urn. Our background knowledge ensures that one of the two following situations holds: Either there is an equal number of white and black balls in the urn (H_1) or there are 75% white balls and 25% black balls in the urn (H_2). Then, an 'independent and identically distributed trial' is defined as a trial where the result of the first draw has no impact on the result of any other draw, given one of the statistical hypotheses (i.e. either H_1 or H_2). For instance, conditionalizing on E_2 does not alter the probability of E_1 given H_1 ($P(E_2|H_1.K) = P(E_2|H_1.E_1.K)$), or in more colloquial terms, H_1 *screens off* E_1 from E_2.[15] Therefore, we describe evidential diversity in terms of probabilistic screening-off.

Now we can state the condition that if evidential diversity (=screening-off) is warranted, then the degree of support lent by the two pieces of evidence should exceed the degree of support lent by any single piece:

> (D)[16] If E_1 and E_2 individually confirm H relative to K and if H screens off E_1 from E_2 (i.e. if (5.9) is satisfied) then for an adequate measure of confirmation $\mathfrak{c}$,[17]
>
> $$\mathfrak{c}(H, E_1.E_2, K) > \mathfrak{c}(H, E_1, K). \tag{5.10}$$

Actually, (D) makes quite mild claims: nothing is said about the *degree* to which adding a piece of independent evidence raises the degree of confirmation. It is just claimed that an additional piece of confirming evidence which is in some sense independent of the original piece of evidence raises the degree of support at all. Since this requirement is so mild, all reasonable measures of confirmation should satisfy (D).

Unfortunately, I did not manage to find analytic results (i.e. proofs or countermodels) for all six measures with respect to (D). Countermodels have

[15] This corresponds, by the way, to independence *conditional on a common cause.* Reichenbach (1956) discusses such 'conjunctive forks' in detail.

[16] The condition (D) is the condition (D') taken from Fitelson 2001a, footnote 67.

[17] Of course, due to the symmetry between E_1 and E_2, equation (5.10) entails that $\mathfrak{c}(H, E_1.E_2, K) > \mathfrak{c}(H, E_2, K)$.

been found for s and c (Fitelson 2001a), and it is obvious that l satisfies (D), but for d, r and z, the issue is more complicated. Computer searches indicate that they 'probably' satisfy (D), but no rigorous proofs have been found so far. The following fact shows that the behavior of d, r and z with respect to (D) is closely tied to each other.

Fact 5.2 *The following four claims are equivalent:*

- *Confirmation measure d satisfies (D).*
- *Confirmation measure r satisfies (D).*
- *Confirmation measure z satisfies (D).*
- *If the presuppositions of (D) are satisfied, then $P(H|E_1.E_2.K) > P(H|E_1.K)$.*

Proof: It is obvious that the first three conditions are all equivalent to the fourth, e.g. for d, we have

$$\begin{aligned} & d(H, E_1.E_2, K) > d(H, E_1, K) \\ \Leftrightarrow\ & P(H|E_1.E_2.K) - P(H|K) > P(H|E_1.K) - P(H|K). \end{aligned}$$

Analogously for r and z.□

Remark: Amending the presuppositions of (D) with the requirement that E_1 and E_2 be *unconditionally independent* of each other ($P(E_1.E_2|K) = P(E_1|K)\,P(E_2|K)$) yields satisfaction of (D) for d, r and z.

Hence, either d, r and z all satisfy (D), or they all violate (D). Since no countermodels have been found, even with the help of computer programs, we conjecture that they all satisfy (D). At least, in all existing examples these three measures recognize the confirmational power of independent evidence.

Now we can summarize our results and compare the various measures of support. To this end, let us have a look at table 5.1. Only three measures have survived the test of seven uncontentious adequacy constraints: d, l (plus its ordinal equivalent k) and z. All other measures violate at least two constraints and drop out of the picture. The next section examines the three remaining measures in detail.

Measure of Support — Adequacy Condition	d	r	k, l	s	c	z
Surprising Evidence	✓	✓	✓	✓	✓	✓
Irrelevant Conjunctions	✓	×	✓	✓	✓	✓
Evidential Diversity (D)	✓?	✓?	✓	×	×	✓?
Commutative Symmetry (CS)	✓	×	✓	✓	×	✓
Evidence Symmetry (ES)	✓	✓	✓	×	×	✓
Strong Law of Likelihood (SLL)	✓	×	✓	✓	✓	✓
Weak Law of Likelihood (exclusive version)	✓	✓	✓	✓	✓	✓

Table 5.1: An overview over the adequacy constraints on confirmation measures.

5.3 A Bayesian account of evidential favoring

Measuring inductive support on a Bayesian account has been criticized as being too subjective for scientific purposes. Indeed, all three measures that have survived the scrutiny of the previous section exhibit subjective elements. Usually, the likelihoods of the evidence under the hypothesis $P(E|H.K)$ are 'objective', e.g. H posits a specific statistical distribution and E describes specific events which naturally have a definite probability given H. For example, if H asserts that a specific coin is fair, then the probability of observing 'heads' in three independent and consecutive tosses of that coin (E) is equal to 1/8. Examples of this kind abound in the statistical literature. But the degree of support depends on $P(H|K)$ or $P(E|\neg H.K)$, too, and unless one is an objective Bayesian, those probabilities are (partially) open to subjective choice, affecting the degree of support E lends to H. This is an outspokenly subjective component in Bayesian confirmation theory and most obvious for d and z since both measures are based on the difference between the a-priori-credence and the a-posteriori-credence. But also for l, the calculation of $P(E|\neg H.K)$ is far from trivial. Admittedly, if there are only two hypotheses H_1 and H_2 which impose definite likelihoods on E, then $P(E|\neg H_1.K) = P(E|H_2.K)$, and the problem vanishes because the latter term is fixed. But often, there are more than two hypotheses at stake – e.g. we have three hypotheses, H_1, H_2 and H_3 – and then, the computation of

$P(E|\neg H_1.K)$ is much less straightforward:

$$P(E|\neg H_1.K) = \frac{1}{1 - P(H_1|K)} \left[P(E|H_2.K)P(H_2|K) + P(E|H_3.K)P(H_3|K) \right] \tag{5.11}$$

so that prior probabilities come in again. This is also called the *problem of the catch-alls*: first, the distribution of the prior probabilities is required for computing (5.11), second, H_3 may be an unspecified 'catch-all-hypothesis', summing up all situations where neither H_1 nor H_2 is the case. Then, even the calculation of $P(E|H_3.K)$ may be difficult because no finite decomposition may exist.

Since measures of support depend on subjective input, Bayesian reasoning is often believed to be too subjective for scientific purposes (Mayo 1996, Royall 1997, 9-11). Instead, Richard Royall (1997) suggests to turn one's attention to relations of *evidential favoring*, i.e. to claims of the sort that evidence E favors a certain hypothesis H_1 over another one H_2, relative to given background assumptions K. To Royall's mind, these relations are crucial for inductive inference in science – theory testing is always *contrastive*, and observations are never evidence for a hypothesis *simpliciter*, but only compared to the available rivals. There are just no practical cases where we test hypotheses without having at least some vague alternatives in mind. Therefore Royall suggests the following criterion:

> (LL): E favors H_1 over H_2 relative to K if and only if ('$\Leftrightarrow$') $P(E|H_1.K) > P(E|H_2.K)$, and the strength of evidence for H_1 over H_2 is measured by $P(E|H_1.K)/P(E|H_2.K)$.[18]

Indeed, that criterion avoids the cumbersome computation of prior probabilities and the catch-all problem: the strength of the evidence is merely the likelihood ratio of the competing hypotheses. The reader will have noticed the strong analogy to the SLL, but here we are concerned with relations of favoring, not of evidential support. So the objections to the SLL do not directly apply. For a Bayesian, however, it is natural to reduce relations of evidential favoring to relations of inductive support: the hypothesis that is better confirmed is favored over its rival, too.

> (B): E favors H_1 over H_2 relative to K if and only if evidence E confirms H_1 better than H_2 with respect to an adequate measure of support $\mathfrak{c}$, i.e. $\mathfrak{c}(H_1, E, K) > \mathfrak{c}(H_2, E, K)$.

[18] See Royall 1997, 2 and Hacking 1965, chapter V.

Again, it can be objected that the theory of evidential favoring stated in (B) is too subjective: reference to prior probabilities or catch-all likelihoods is inevitable, given what we have found out about measure of support. But (LL) is vulnerable to serious criticism, too: Royall's (1997) 'likelihoodist' theory of evidential favoring (LL) can be represented as a special Bayesian theory of favoring, using the log-ratio measure r![19] Thus, his criticism that a Bayesian cannot quantify scientific evidence in an objective manner does not seem to be justified because his own theory is just a variation of a Bayesian account of evidential favoring.

It would be premature, though, to argue that r's inadequacy as a measure of inductive *support* disqualifies it for determining relations of evidential *favoring*. Recall the problem of irrelevant conjunctions. I do not find it odd to say that E confirms H more than $H.X$ (let X be an irrelevant conjunct) while holding at the same time that E does not favor H over $H.X$ (since H does not predict E any better than $H.X$). It is not so clear that support is so intimately tied to evidential favoring as (B) insinuates: inductive support measures strength of an inductive argument or increase in credibility. But if an improbable and a probable hypothesis confer the same likelihood on the evidence, the evidence seems to be neutral between the two, although it can have a different *relevance* for the two hypotheses, in the sense that their probability is more or less significantly changed. Therefore (LL) cannot be ruled out on the grounds that it relies on an inadequate measure of inductive support.[20] When we disentangle support and favoring, it is not clear why the representation of (LL) as a Bayesian theory of favoring should trouble a likelihoodist. She could just argue that favoring and support measure very different things and that a Bayesian representation of (LL) is just a representation and essentially a mathematical gimmick. Furthermore, (LL)'s simplicity and its independence of prior probabilities are very attractive features. In a sequel paper, Royall (2000) presents many fruitful applications of an (LL)-based approach to statistical evidence (more on this in chapter 7) and develops it towards a full theory of inductive reasoning. Thus, the debate between (LL) and the various forms of (B) remains open. The charge of subjectivity raised against Bayesian confirmation theory is certainly misguided with respect to measures of support (most forms of Bayesianism are intended as a subjective theory), but it is equally clear that subjectivity is

[19] This was noted by Branden Fitelson (2007).

[20] Fitelson seems to make such an argument against r in footnote 22 of his 2007.

less desirable in relations of evidential favoring (e.g. when testing theories against each other and quantifying evidence).

I do not want to decide the debate, but before concluding, I would like to point out some features of the various versions of (B). First, we could use the log-likelihood measure l which is also a viable measure of support. Assume that there are three mutually exclusive hypotheses (H_1, H_2 and H_3) and that H_1 and H_2 are confirmed by E, but that $P(E|H_1.K) = P(E|H_2.K)$. Then, the circumstances when l favors H_1 over H_2 can be easily described:

$$\begin{aligned} l(H_1, H_2, E, K) &= \frac{\log P(E|H_1.K) - \log P(E|\neg H_1.K)}{\log P(E|H_2.K) - \log P(E|\neg H_2.K)} \\ &> 1 \text{ iff } P(E|\neg H_1.K) < P(E|\neg H_2.K). \end{aligned}$$

Under the given assumptions it is not difficult to see that

$$\begin{aligned} l(H_1, H_2, E, K) > 1 &\Leftrightarrow P(E|\neg H_1.K) < P(E|\neg H_2.K) \\ &\Leftrightarrow P(H_1|K) > P(H_2|K). \end{aligned}$$

In other words, if two mutually exclusive hypotheses confirmed by E confer equal likelihood on E, then l always favors the hypothesis which had the higher prior probability beforehand. Again – it may be perfectly fine that a probable hypothesis H_1 is supported to a higher degree because the change in the credence in H_1 will be much more pronounced than the change in the credence in H_2.[21] This does not imply, however, that the evidence *favors* H_1 over H_2 since both hypotheses are equally good predictors of the observed event. The property of l to always favor the more probable hypothesis in the case of equal likelihoods is somewhat awkward for a measure of evidential favoring.

Finally, relations of evidential favoring are often blurred by the fact that hypotheses share common content. In principle, we would like to have an account of favoring that is purely contrastive: it would be nice to have a measure that reduces evidential favoring relations to evidential favoring relations between the proper (non-intersecting) part of the hypotheses. Among the measures which are still in the game, d is the only measure that satisfies that constraint. First, we note that d is *additive*:

[21] This is an instance of the *Matthew effect* in Bayesian confirmation theory.

Fact 5.3 *Let H_1 and H_2 be mutually exclusive, i.e. $H_1 \models \neg H_2$. The difference measure d fulfils*

$$d(H_1 \vee H_2, E, K) = d(H_1, E, K) + d(H_2, E, K). \tag{5.12}$$

Proof: trivial.

Now, let H_1 and H_2 be two overlapping hypotheses. First we decompose H_1 and H_2 suitably, then we use the additivity of d to 'cut out' the overlapping part:

$$\begin{aligned}
& \text{`}E\ d\text{-favors } H_1 \text{ over } H_2 \text{ relative to } K\text{'} \\
\Leftrightarrow\quad & d(H_1, E, K) > d(H_2, E, K) \\
\Leftrightarrow\quad & d(H_1.\neg H_2, E, K) + d(H_1.H_2, E, K) > d(\neg H_1.H_2, E, K) + d(H_1.H_2, E, K) \\
\Leftrightarrow\quad & d(H_1.\neg H_2, E, K) > d(\neg H_1.H_2, E, K). \qquad (5.13)
\end{aligned}$$

The two hypotheses in the last line are evidently disjoint. The calculation shows that, in order to evaluate favoring relations, it is sufficient to look at the parts of H_1 and H_2 *which exceed the mutual overlap.* According to d, favoring relations between disjoint hypotheses determine all favoring relations. So hypothesis testing is again a *contrastive* activity: Only those parts of a hypothesis that are incompatible with the rival hypothesis are compared to each other. All favoring relations are reduced to favoring relations between mutually exclusive hypotheses which is convenient both from a systematical and a practical point of view. Since l and z are not additive in the sense of (5.12), they do not yield a similar result.

In this brief section, I have argued that the problems of inductive support and evidential favoring are not co-extensional. In particular, not all good measures of support are good measures of evidential favoring, and vice versa. Indeed, the most attractive 'likelihoodist' account of evidential favoring (LL) can also be described as a Bayesian account of favoring based on an inadequate measure of support. I have tried to dispel the worries associated with that representation and discussed some properties of Bayesian theories of evidential favoring.

5.4 Logicality and the problem of old evidence

A principled decision in the discussion of measure of support concerns the question how seriously one should take the analogy to logical entailment. Two projects in finding a measure of support interfere: on the one hand, we aim at a quantitative generalization of logical entailment in the sense of an inductive logic, i.e. we would like to quantify the strength of the inductive argument E gives for H. This is a *structural* approach to Bayesian confirmation theory. On the other hand, we would like to find an explication of support and confirmation that is suitable for application to scientific examples and that bears some resemblance to our ordinary concept of confirmation.

The first project leads us to the logicality condition:

> **Logicality (L):** For an adequate measure of confirmation $\mathfrak{c}$, $\mathfrak{c}(H, E, K)$ takes its maximal [minimal] value if ('$\Leftarrow$') $E.K \models H$ $[E.K \models \neg H]$.[22] [23]

In other words, logicality demands that a measure of inductive support generalizes deductive entailment. The idea behind (L) seems to be that a measure of support should quantify the power of an inductive argument, and clearly, conclusive, deductive arguments (i.e. arguments that leave no room for doubt) are the strongest conceivable arguments. Maximal support corresponds to maximal strength of argument, hence $\mathfrak{c}$ should be maximal in case of $E.K \models H$.

Which measures of support satisfy (L)? If $E.K$ entails H, we get

$$P(H|E.K) = 1 \qquad\qquad P(E|\neg H.K) = 0.$$

Hence,

$$k(H, E, K) = \frac{P(E \mid H.K) - P(E \mid \neg H.K)}{P(E \mid H.K) + P(E \mid \neg H.K)} = 1 \qquad \text{if } E.K \models H. \tag{5.14}$$

[22] See Fitelson 2001a, 2006.

[23] Notably, there is a related condition (Ex1) suggested by Crupi et al. (2007, 232): measures of support have to respect the ordinal structure of logical entailment, i.e. if $E_1.K \models H$, then E_1 confers greater support to H than any piece of evidence that does not entail H (jointly with K), etc. This condition is neither necessary nor sufficient for logicality – the task of finding counterexamples is left to the reader. But in practice, all measures of confirmation that satisfy logicality satisfy (Ex1), and vice versa. Therefore the discussion of (Ex1) boils down to a discussion of logicality.

From (5.14), it is clear that k fulfils (L) since it can only take values between -1 and 1. Due to the ordinal equivalence of k and l, the satisfaction of (L) transfers to l, too. An identical result can be shown for z. However, d does not fulfil (L): If $P(H|K)$ is already quite large, $d(H, E, K)$ will not be maximal if $E.K \models H$. $d(H, E, K)$ would be greater if H were a quite improbable hypothesis significantly boosted, but not conclusively supported by $E.K$.[24] Fitelson judges the disagreement in favor of l. He writes:

> "I take it as intuitively clear that the strength of the support E provides for H in this case should *not* depend on how probable H is [a priori, J.S.]. [...] After all, evidential support is supposed to be a measure of how strong the evidential *relationship* between E and H is, and deductive entailment is the strongest that such a relationship can possibly get."[25]

This passage again insinuates the 'strength of inductive argument' explication for a measure of support. In particular, if E deductively entails H, the prior probability of H should be irrelevant for the degree of evidential relevance. So Fitelson's quote provides an argument against d and in favor of l and z.

I believe, however, that both (L) and Fitelson's argument against dependence on prior probabilities can be doubted in a reasonable way. We are interested in the evidential relevance of E *for* H. Hence, we would like to know *the effect* of the evidence on H. But for an estimation of the effects of E on H, it is very reasonable to look at the priors. This is unanimously accepted – for instance, the WLL, defended by Fitelson himself, entails that, if the likelihoods of the evidence are (roughly) equal under two competing hypotheses, the prior probabilities will often decide which hypothesis is better supported (see the previous section). So it is not clear why this kind of reasoning should be suspended in the case of $E.K \models H$. Imagine an apt chess player P who has a completely winning position. His confidence in the hypothesis H : 'I (P) will win the game' will be very close to 1.[26] Then evidence E_1 occurs – the opponent resigns. E_1 implies H (together with some innocent background knowledge), so E_1 confirms H maximally according to (L). Nonetheless we are inclined to believe that P would have won the game

[24] Crupi et al. (2007) show that among an exhaustive set of candidate measures of support, only l and z satisfy the closely related condition (Ex1), see the previous footnote.

[25] Fitelson 2001a, 42, original emphasis.

[26] It is not equal to 1 since the player knows that he might still blunder, suffer a heart attack or fail to win the game for another unusual reason.

anyway, so the evidence E_1 was not very *informative.*[27] We can pursue the issue further: In his next game P has a dreadful position, so his confidence in H is quite low. By a matter of accident, his opponent forgets to press the clock after making his move and loses on time (E_2) – P wins the game. Due to the different point of departure, event E_2 seems to be much more relevant for the result of the game than event E_1. For any practical purposes, the information given by E_2 is much more valuable and important than the information given by E_1 because our judgments drastically shift. Moreover, *ceteris paribus*, surprising evidence should confirm to a stronger degree than expected evidence, as pointed out previously. In the above example, E_2 was much more surprising than E_1, and this should be mirrored in the measures of confirmation. However, l and z are unable to distinguish both cases, whereas d assigns different values to the first and the second case. Thus, d accounts for the informational richness and surprise inherent in E_2 when compared to E_1. Evidential relevance is not only logical strength of argument, but is also mirrored in the changes of our epistemic attitudes towards the hypothesis.[28]

This has a more general implication: If we subscribe to Fitelson's view, we are drifting away from relative confirmation to absolute confirmation. According to l, z and ordinally equivalent measures, E is maximally relevant evidence for H if and only if ('$\Leftrightarrow$') $P(H|E.K) = 1$. *This is characteristic of measures of absolute confirmation and not in the spirit of relative confirmation.* In particular, the value of $P(H|K)$ (e.g. whether it is close to 0 or 1) does not matter at all. However, for d, neither direction of the above equivalence holds. These points speak in favor of d and against its competitors l and z. Logicality (L) presupposes the very claim that is at stake – that degree of support is a quantitative generalization of deductive strength of argument to inductive arguments.

The difference between the two views of relative confirmation – strength of inductive argument and increase in credibility – is also crucial in the discussion of a very longstanding problem of Bayesian confirmation theory: the *problem of old evidence.* The problem itself is very venerable, but the debate between various resolution proposals has been blurred by the lack of distinction between these two senses of confirmation which is incorporated

[27] In identifying a-posteriori plausibility and informativeness as confirmation-conducive factors, Huber (2005) makes a similar point in different words.

[28] Of course, I do not doubt that E confirms H if E entails H. I only doubt that this is a sufficient criterion for confirmation to a maximal degree.

in the measure d on the one side and the measures l and z on the other side. But what is the problem itself? Traditionally, the context of discovery of Einstein's General Theory of Relativity (GTR) is used to explain it. In the nineteenth century, it was noticed that the perihelion of the planet Mercury was advancing, contrary to the model of Newtonian mechanics whose predictions could not explain the entire precession effect. People tried to explain the perihelion advance by the motion of the other planets, but those attempts were not successful. Therefore astronomers and physicists were continually worrying about Mercury's anomalous perihelion precession which was much larger than what experimental error could account for.[29] For instance, the French mathematician Urbain Le Verrier postulated a further, undiscovered planet named 'Vulcan' inside the Mercury orbit. Others ascribed the advance to a slight oblateness of the sun. Such *ad hoc* hypotheses failed to gain empirical corroboration and were eventually rejected. In the early twentieth century, Albert Einstein tried hard to find a theory that was compatible with the Mercury's anomalous perihelion advance. In November 1915, he wrote down the final version of the General Theory of Relativity which was able to explain the perihelion advance by taking into account the curvature of spacetime. It was generally appreciated that the theory was able to resolve this longstanding riddle of celestial mechanics. Earman (1992, 119) motivates the significance of the old evidence problem by claiming that physicists assigned a higher confirmatory value to the perihelion advance than to other classical tests of GTR – the bending of light (demonstrated by Eddington in 1919) and the gravitational redshift (proven in the Pound-Rebka experiment in 1959). This is, however, a disputable claim, e.g. Wiechert (1920) outrightly claims the converse.[30]

Be this as it may, Bayesians struggle with modeling old evidence. The Mercury perihelion advance (E) was 'old news' at the time GTR was formulated, i.e. it was a part of the available background knowledge: $P(E|K) = 1$. Thus, E could not confirm GTR according to Bayesian conditionalization.[31]

[29] The difference between the predictions and the actual advance was about 43 seconds of arc per century.

[30] "Die Sachlage für die Relativitätstheorie ist hier [bei der 1919 nachgewiesenen Lichtablenkung] noch um vieles günstiger als bei der Perihelbewegung des Merkur, welche lange bekannt war und für welche vielfach Erklärungen gegeben waren." (Wiechert 1920, 301)

[31] Note that the problem would *not* vanish but only be slightly mitigated if E had been known with 'approximate certainty', e.g. $P(E|K) = 0.9999$.

The problem may actually be relabeled as the 'problem of introducing new theories' since the fact that E was observed before GTR was formulated was the cause of so many Bayesian headaches. In a similar vein, we can also ask whether hypotheses are confirmed by data they were constructed to explain – Einstein definitely aimed at a theory that could account for the Mercury perihelion.

The Bayesian's problem seems to be that in principle, all present, past and future theories are included in her partition of possible theories and hypotheses. In an idealized Bayesian picture, all those theories have been formulated at the beginning of time and the rational degrees of belief which we assign to them are simultaneously updated when new information is coming in. But clearly, this is not the way science works.

Garber (1983), Jeffrey (1983) and Niiniluoto (1983) want to dissolve the dilemma by means of relaxing the unrealistic condition that Bayesian agents are logically omniscient. They would like to model the learning of logical truths. To motivate their approach, note that the old evidence problem occurred because at the time the perihelion advance E was discovered, GTR was not yet formulated and thus not present in the agents' belief states. Therefore GTR could not be confirmed by E before E became 'old evidence'. Garber, Jeffrey and Niiniluoto make the following diagnosis: It was not the *formulation* of GTR that led to a confirmation of GTR by E – rather it was Einstein's discovery that GTR *entailed* E which spoke so much in favor of GTR. By allowing for the learning of logical truths and relaxing the condition of logical omniscience that is imposed on an ideal Bayesian agent, Garber, Jeffrey and Niniiluoto purport to dissolve the problem of old evidence. To this end, they enrich the object language by sentences of the form $H \vdash E$ which are assigned degrees of belief between zero and one, just as all other sentences. This provides a framework for updating one's belief on such sentences. In particular, it may happen that

$$P(H|(H \vdash E).E.K) > P(H|E.K). \tag{5.15}$$

Thereby the Garber/Jeffrey/Niiniluoto approach is able to model that the confidence in GTR was substantially raised by the discovery that it explained the Mercury perihelion advance, although the perihelion data themselves were old news (see equation 5.15). We then see how the introduction of GTR and the discovery that GTR $\vdash E$ raised the credibility of GTR and confirmed it in the Bayesian sense. But it can rightfully be asked which question was actually answered by their approach. One question was whether *Mercury's perihelion data* confirmed GTR in 1915, given Einstein's background knowledge. This seems to be a question about the structural relationship between E and GTR. Garber, Jeffrey and Niiniluoto instead answer the question whether learning the logical truth GTR $\vdash E$ increased Einstein's *confidence* in GTR. These questions are, however, not equivalent, as the preceding discussion of logicality makes clear. Modern physics textbooks accompany the introduction of GTR with the derivation that GTR explained the Mercury perihelion. For contemporary physics students who learn GTR, there is no time point when $P(\text{GTR} \vdash E|K) < 1$, but still they want to say that E confirmed GTR.[32] The Garber/Jeffrey/Niiniluoto approach captures the 'increase in degree of belief' rationale of Bayesian confirmation theory, but it does not explicate the structural relation of inductive support between GTR and E. This is the same discussion we already encountered when discussing logicality, and it can be extended to a very principled criticism of Bayesian confirmation theory, raised by Clark Glymour:

> "[...] that relation [of evidential relevance, J.S.] depends somehow on structural, objective features connecting statements of evidence and statements of theory. [...] There must be relations between evidence and hypotheses that are important to scientific argument and to confirmation but to which the Bayesian scheme has not yet penetrated."[33]

We have already made the acquaintance of Glymour's own answer: he abandons probabilistic theories in favor of the qualitative bootstrap approach which is, to his mind, able to capture the logic of scientific confirmation.[34] But I do not think that Glymour's conclusion is inevitable. To see this, I would like to have a look at another proposal to resolve the problem.

Howson and Urbach (1993) suggest to modify the background knowledge on which we conditionalize: instead of conditioning on the totality of the background knowledge at the time GTR was introduced, we should omit the old evidence E (i.e. the perihelion data) from the background knowledge. Then we can measure the support for GTR by "the extent to which the addition of E to the remainder of what you currently take for granted, would

[32] See Earman 1992, 130.

[33] Glymour 1980a, 93.

[34] See chapter 3 for discussion.

cause a change in your degree of belief in [GTR]."[35] So instead of measuring a change in our *actual* belief state, we measure a change in a *counterfactual belief state* where the old evidence and all other pieces of confirming or disconfirming evidence are not yet known. Instead of the background assumptions K we have to work with $K - \{E\}$.

This kind of resolution has attracted much criticism. Replacing K by $K - \{E\}$ sounds easy, but in fact it is highly problematic how to construct such a consistent and realistic belief function that eliminates E. What if there are logical dependencies between E and other parts of K? Is E really separable from K? What about the description of the experimental setup used to generate E? The need to eliminate E (and all data that are closely related to E) from the background knowledge conflicts with enormous practical difficulties:

> "We cannot merely throw out E and whatever entails E out of the body of accepted beliefs; we need some rule of determining a counterfactual degree of belief in E."[36]

Indeed, it is unclear how such a rule could ever be extracted from practice. This is not surprising: such a rule would have to elicit the likelihoods of E under the competing hypotheses as well as the priors of the hypotheses in the counterfactual belief state. This objection is sound and shows that much work has yet to be done in order to cope with the problem of old evidence. But this does not entail that the idea to work with counterfactual degrees of belief is entirely hopeless. Assume for reasons of simplicity that we have only two competing hypotheses, H and $\neg H$. Let the likelihoods of the evidence under the competing hypotheses be objectively determined probabilities (which is often plausible in the case of statistical hypotheses, see chapter 4). Then likelihood-based measures as $l(H, E, K) = \log[P(E|H.K)/P(E|\neg H.K)]$ are able to measure the support E lends to H without taking recourse to the fact that E was in some sense old evidence. True, we have to eliminate E from the background knowledge in order to avoid trivialization, but sometimes, this may work. In many cases, e.g. when E is entailed by the hypothesis or when H and $\neg H$ are statistical hypotheses assigning definite probability values to $P(E|H.K)$ and $P(E|\neg H.K)$, such a calculation will be straightforward. For other measures, this need not be the case, e.g. conducting such a

[35] Howson and Urbach 1993, 271, notation of symbols changed.

[36] Glymour 1980a, 87.

procedure with the difference measure $d(H, E, K) = P(H|E.K) - P(H|K)$ would require a cumbersome elicitation of prior probabilities (similar for z).

Hence, I recommend a double-tracked Bayesian strategy, corresponding to the ambiguity of Bayesian confirmation and the different Bayesian explications of degree of support which also create different versions of the problem of old evidence. If we understand support as increase in credibility (in the sense of d or related measures as r), the problem can be tackled by means of the Garber/Jeffrey/Niiniluoto route and the explicit incorporation of logical learning. This strategy is misguided, however, if confirmation is identified with the strength of an inductive argument (quite similar to qualitative confirmation, by the way). Here, the Howson/Urbach approach to work with counterfactual belief function may work, and deviating from Howson and Urbach, I believe that measures as l are best suited for this task. Thus, the criticism of Bayesian solutions to the problem of old evidence and Bayesianism in general is not always fair because it is not always clear which kind of inductive support is the target of the problem. Therefore, all answers can only be partially successful. Surely, the present accounts of the problem are far from being complete, and their empirical success is uncertain. No single approach can solve all aspects of the old evidence problem, but several approaches can tackle different varieties of the problem.

An interesting variation of the problem of old evidence is discussed in Fitelson (2001a, 2001b). Here, the question is discussed in how far knowing previously observed evidence affects the degree of support inherent in later evidence. We deal with the following model which is encapsulated in our background knowledge K:

> "An urn has been selected at random from a collection of urns. Each urn contains some balls. In some of the urns the proportion of white balls to other balls is x and in all other urns the proportion of the white balls is y, $0 < x$, $y < 1$. The proportion of urns of the first type is z, $0 < z < 1$. Balls are to be drawn randomly from the selected urn, with replacement."[37]

Now, select an urn U and let H be the hypothesis that the proportion of white balls in U is x. Let W_i state that the i-th draw from U is a white ball. Then, Fitelson proposes the following condition:

[37] Fitelson 2001b, 128-29.

> *Urn Condition (UC)*: For all adequate measures of confirmation $\mathfrak{c}$ and all urn models, regardless of the values of x, y and z:
>
> $$\begin{aligned}\mathfrak{c}(H, W_1, K.W_2) &= \mathfrak{c}(H, W_1, K)\\ \mathfrak{c}(H, W_2, K.W_1) &= \mathfrak{c}(H, W_2, K).\end{aligned} \tag{5.16}$$

We realize that l satisfies (5.16) whereas d and z violate (5.16). – Accepting *(UC)* amounts to saying that the (known) result of W_1 does not influence the evidential relevance of W_2 for H – and vice versa. We obtain the same evidential relevance like in a situation where the result of W_1 is indeterminate. The idea is that the single draws are independent of each other – no result of a draw has an impact on the evidential relevance relation between H and the result of another draw. More generally, H screens off W_1 and W_2 relative to K, so knowing one of the results should not affect the evidential impact of the other result $[c(H, W_2, K.W_1) = c(H, W_2, K)]$. Fitelson thus sharpens *(UC)* into the following condition

> *Screening-Off Condition (SC)*: If $\mathfrak{c}$ is an adequate measures of confirmation and H screens off E_1 from E_2, then
>
> $$\begin{aligned}\mathfrak{c}(H, E_1, K.E_2) &= \mathfrak{c}(H, E_1, K)\\ \mathfrak{c}(H, E_2, K.E_1) &= \mathfrak{c}(H, E_2, K).^{38}\end{aligned} \tag{5.17}$$

Now, let us come back to the urn model. Before conducting the actual experiment, assume that a considerable number of balls has already been drawn from the urn, all of them being white. Recall that H was the hypothesis that the proportion of white balls in the urn is equal to x and assume furthermore that $x > y$. Then we draw three white balls out of the urn (E). Naturally we hold that E, *given that the first balls were all white*, will not confirm H to the same degree as in a situation where no information about other draws was available. If we know that the first N balls are white, we assign H such a high credibility that further confirmation – in the sense of increase in credibility – will be negligible. Three balls more do not seem to make such a huge difference. However, this conflicts with the (SC)-intuition that each draw of a white ball bears the same structural relation to H and should have the same confirming power. Actually, we have discovered a sister problem of the old evidence problem – the problem of *probable hypotheses.*[39]

[39] Christensen 1999, 448-49.

If H is already very likely, there is little room for further support in the sense of increase in credibility. Again, the fundamental ambiguity of confirmation and support becomes evident: On the one hand, there is the reading of confirmation as increase in credibility, on the other hand, there is the reading of confirmation as the strength of an inductive argument. The dissent on (SC) and (UC) exemplifies this ambiguity again. It seems to me that both readings have to be preserved: Certainly, structural, belief-state-independent relations of evidential support are important for science, as argued by Glymour. This corresponds to the intuition behind (SC), too, and the idea that confirmation generalizes logical entailment. On the other hand, that reading does no longer measure the evidential relevance of E – the impact which the evidence exerts on the epistemic status of the hypothesis.

5.5 Summary

This long chapter has presented the main idea of Bayesian confirmation – increase in degree of belief – and discussed various measures of support. A set of mild adequacy rules out half of our proposed measures (various other measures that were not discussed here also fail to satisfy those conditions). The remaining measures mitigate the problem of irrelevant conjunctions, account for the power of surprising evidence and evidential diversity and do not exhibit vicious symmetry properties. Furthermore they satisfy the WLL, vindicating the significance of successful prediction for inductive support.

At the end of the day, the difference measure d, the log-likelihood measure l (plus its ordinal equivalent, the Kemeny-Oppenheim measure) and Crupi's and Tentori's measure z are left in the basket. We have elaborated that the difference measure d, although presently not fashionable, can be successfully defended against attempts to prove its inferiority, e.g. in a Bayesian theory of evidential favoring. The discussion of logicality (L) has revealed that the three remaining measures of confirmation explicate at least two different senses of inductive support: evidential relevance for the hypothesis and increase in credibility (d) and strength of an inductive argument and generalization of deductive entailment (l, z). Among the latter measures, l is often easier to calculate, but z has the advantage of being the only measure of support whose symmetry properties mirror the symmetry properties of deductive entailment. Ultimately, the context of application decides whether d, l or z should be preferred. This does not transfer to other measures of support, though, since their drawbacks (e.g. certain vicious symmetry properties) are brought to bear in any context of application.

The above distinction allows us to see the problem of old evidence in greater clarity, too, and to be more charitable towards the attempted solutions. Moreover, we note that evidential favoring and degree of support can fall apart and that measures of evidential favoring need not be reduced to Bayesian measures of support. In particular, we have defended the most attractive, widespread and fruitful theory of evidential favoring – Richard Royall's likelihoodism – against Bayesian attacks presupposing that evidential favoring reduces to degree of support. In the remainder of the book, we will contrast the Bayesian approach to other schools of statistical inference.

Chapter 6

Statistical Hypothesis Testing

6.1 Statistics and the sciences

The last chapters have represented confirmation and support as changes in rational credences, building on a probabilistic calculus for those credences. In particular, we have compared various measures of support to each other. So far, everything was applicable not only to statistical, but also to deterministic theories and hypotheses. That was quite reasonable – many historical cases of confirmation in science are located in a deterministic framework (as the GTR/Mercury example discussed in the last chapter), and a confirmation theory should cover those cases, too. But now, I would like to focus on confirmation in statistics – the science pertaining to the collection, analysis and interpretation of data and the probabilistic explanation of observed events. To fully understand the significance of statistical reasoning in confirmation theory, some historical remarks may help.

The application of probabilistic models and statistical regularities emerged in the natural sciences, as witnessed by famous applications like Gregor Mendel's laws of inheritance. Pioneer statisticians were often based in the natural sciences, e.g. Ronald A. Fisher who expanded and refined Mendel's theory was not only an important mathematician, but also a leading geneticist. The benefits of using probabilistic models in quantifying risk and uncertainty were soon acknowledged, and nowadays, statistical reasoning is the most everyday activity in the natural sciences. Many physical processes are so complex and hard to understand that non-probabilistic models fail – the related disciplines of geophysics, meteorology and climate science provide the standard examples. Only models that explicitly account for the uncertainty about the underlying physical processes are descriptively adequate. In any case, it is hard to imagine modern science without statistical methods – even quite remote and formerly non-mathematical disciplines as ecology have nowadays been invaded by statistical methods.[1] Probabilistic models often have instrumental value – they are not adopted because the stochastic regularities are believed to give a realistic picture of the world or to be fundamental laws of nature. Instead, they are adopted simply because they make the best predictions. Thus, we are not interested in the matter that is uncertain (e.g. meteorological processes), but in modeling uncertainty and making predictions on the basis of uncertainty models.[2]

It took some more time to bring formal methods to the social sciences which have recently undergone a strong formalization and mathematization. At the beginning of the 20th century, the use of mathematical methods in the social sciences was still in its early stages. The positivist program and the rise of falsificationist methodology in philosophy of science had, however, a deep impact on the social sciences. Abstract theorizing was more and more replaced by experiment- and observation-based reasoning that focused on the *observability* of the relevant quantities. The behaviorist program in psychology gives a salient example. This development laid the foundations for the rise of mathematical methods. In particular, statistical analysis of observed data and mathematical modeling came into focus. In the middle of the 20th century, the economist L. J. Savage even came up with an axiomatic theory of rational behavior, the famous expected utility theory, and at a similar time, John von Neumann developed the foundations of game theory which is now a major branch of economics. Such highly abstract and mathematically non-trivial theories yielded a number of rewarding results and opened the way to fruitful research programs. Nowadays, mathematical methods are nearly omnipresent in the social sciences, and even applied sciences (such as business studies) use mathematical tools as decision trees.

The success story of mathematics in the social sciences would be incomplete, however, without explicit consideration of the achievements of statistics. Statistics emerged both from mathematics and the empirical sciences at the beginning of the 20th century and proved to be an indispensable tool in data analysis and inference from data to general conclusions. Statistical methods are used to discover causal dependencies and to build and to assess

[1] See the contributions in Taper and Lele 2004.

[2] See Lindley 2000.

mathematical models of real-world processes – take, for instance, goodness-of-fit tests as the famous χ^2-test or contrived model selection procedures. Econometrics perspicuously illuminates the coherence of mathematical modeling and statistical data analysis, e.g. in the analysis of time series. Most statistical analysis is performed in the framework of Neyman-Pearson statistics, i.e. statistical inference builds on the reliability of the employed procedures and prohibits the use of subjective probabilities which are prevalent in the alternative Bayesian statistics. We will discuss and contrast both approaches in this chapter.

Statistical methods are, however, easily misunderstood, and many scientists who lack the necessary mathematical sophistication do not know how to interpret them properly. For instance, the use of p-values in the social sciences is such a major source of confusion so much the more as many journals demand that p-values accompany experimental reports. Solving this problem and improving the statistical education of social scientists is actually a major political issue in the social sciences which is reflected in the increasing interest in mathematical training among empirical scientists. Therefore, scrutinizing the foundations of statistics and the proper way to confirm scientific hypotheses is an absolutely essential issue for scientific methodology. This book can, of course, only discuss selected topics, but among them, there will be central topics as the dissent between the various schools of statistical inference, the role of p-values, desiderata for a measure of statistical evidence and the role of experimental design. In the next two chapters, we present the state of art of the debate and argue that Bayesian reasoning provides a unified approach to mathematical modeling and statistical data analysis and is therefore at least a serious alternative to the prevalent Neyman-Pearson school of statistical inference.

6.2 Foundations of Bayesian statistics

The past chapters have introduced the subjective probability interpretation and the Bayesian theory of inductive inference. Naturally, we would like to know how this theory extends to statistical applications and real scientific hypotheses, and whether Bayesian confirmation theory performs well when applied to statistical hypotheses. First, we note that Bayesian confirmation theory can directly be extended to a theory of statistical inference. Bayesian conditionalizers compute posterior probabilities from prior probabilities, akin to the scheme of a probabilistic logic:

$$P(H) = \alpha_1,\ P(E) = \alpha_2,\ P(E|H) = \alpha_3 \quad \models \quad P(H|E) = \frac{\alpha_1 \alpha_3}{\alpha_2}.$$

The same procedure can be applied in a statistical framework where we have a parameter of interest (say, ϑ) whose true value we do not know and random data which depend on ϑ. Then, the competing hypotheses correspond to different values of ϑ and we can assign a prior distribution to ϑ, representing our prior credence in a particular value of ϑ. For instance, ϑ could be distributed according to the standard normal distribution ($\vartheta \sim N(0,1)$) or according to a uniform distribution on a specific interval ($\vartheta \sim U[0,1]$), etc. Then, those prior probabilities are updated on the observed data x to posterior probabilities by means of Bayes's theorem. Assume that $\vartheta \in \mathbb{R}$ (one of the most frequent cases) and that we would like to compute the posterior probability that $\vartheta \in I$, $I \subset \mathbb{R}$. The prior distribution of ϑ is given by the probability density $\phi(\vartheta)$ and the likelihood of x given ϑ is described by $\rho(\vartheta, x)$.[3] Then, the posterior probability of $\vartheta \in I$ is equal to

$$\begin{aligned} P(\vartheta \in I \mid x) &= \frac{P(\vartheta \in I) P(x|\vartheta \in I)}{P(x)} \\ &= \int_I d\vartheta\, \phi(\vartheta) \frac{\int_I d\vartheta\, \phi(\vartheta)\rho(\vartheta, x)}{\int_I d\vartheta\, \phi(\vartheta)} \frac{1}{\int_{\mathbb{R}} d\vartheta\, \phi(\vartheta)\rho(\vartheta, x)} \\ &= \frac{\int_I d\vartheta\, \phi(\vartheta)\rho(\vartheta, x)}{\int_{\mathbb{R}} d\vartheta\, \phi(\vartheta)\rho(\vartheta, x)}. \end{aligned}$$

which can be computed from the probability densities. The main problem of Bayesian inference is the subjective component inherent in the prior probability density $\phi(\vartheta)$ – something that is often deemed inadequate for an 'objective' activity as scientific research. There are some results that try to mitigate the impact of subjective judgments, namely the often-cited *Gaifman-Snir theorem.* Some technical presuppositions set aside, Gaifman and Snir (1982) have shown that the rational credences of two agents who assign credence 0 and 1 to the same events (called *equally dogmatic* agents), will merge and eventually converge against each other as more and more evidence comes in.[4] This results gives a prima facie answer to the charge

[3] Assuming the existence of such densities facilitates the calculations and can often be justified by scientific background theory (Hacking 1965).

[4] See Gaifman and Snir 1982, 208 and Earman 1992, 145-147.

of subjectivity: yes, Bayesian inference is subjective, but as more and more data come in, the opinions of two agents will eventually converge and the initial differences will become negligible. Having a huge set of observations and data compensates for diverging initial opinions. Although the Gaifman-Snir theorem is a major theoretical achievement, their result is not as powerful as it seems at first sight. First, the convergence result is not uniform over the agents' prior probabilities – there is no fixed time point at which the judgments of all equally dogmatic agents with arbitrary prior opinions will have merged. This observation restricts the scope of the theorem. Second, there is in general no bound for the rate of convergence. Sometimes, the Central Limit Theorem (CLT) can do the job and describe the distribution of the (normalized) average of independent and identically distributed random variables. In particular, that distribution becomes more and more skewed around the mean of the random variables, with convergence rate $\sqrt{n}$. But often, we do not have such results at hand, and the Gaifman-Snir does not provide them either. Third and last, scientists often ask the question whether a *particular* observation (say, Eddington's observations of the 1919 eclipse) were evidence for a specific theory (say, GTR). A Bayesian theory of inductive inference has to address this question independent of any considerations about the long run because such long-run observations are not made. Here, the presence of the subjective elements is not at all mitigated by merger-of-opinion results for a long series of observations.

Thus, Bayesian attempts to solve the subjectivity problem have to take recourse to objective Bayesianism or at least to some convention for choosing prior probabilities. This endeavor has, however, a long history of dissent among statisticians. It goes beyond the scope of this work to discuss all proposals that have been made, from Haldane priors (Jaynes 1968) over conjugate priors (i.e. priors which are, given specific likelihood functions, in the same family as the posterior distribution) to predictive priors which are gained from training from the data, and so on. Some ideas to represent informationless or 'ignorant' prior distributions deserve, however, special mention. To bound the effect of the subjective choice of a prior distribution, Bernardo (1979) has suggested to choose the prior distribution that maximizes the expected Kullback-Leibler divergence between prior and posterior distribution, relative to the data. The rational behind this idea is quite the opposite of cross-entropy updating: whereas cross-entropy updating selects the posterior distribution which is closest to the prior distribution, given the information, Bernardo's *reference priors* maximize that distance (and thus assign maximal informational content to the observed data). Jeffreys (1961) has suggested a similarly reasonable criterion, namely to select a prior distribution that is invariant under reparametrization of the parameter space, thus replying to the Bertrand-style paradoxes which we have encountered in chapter 4. All those suggestions are free of subjective elements and constitute objective theories of inductive inference. But the gain in objectivity has its drawbacks, too, since it becomes much harder to incorporate subjective expertise into the statistical inference. This flexibility is one of the attractive features of the personalist version of Bayesianism, and it gets lost when using objective prior distributions. Only objective Bayesianism seems to circumvent the problems of 'objective priors' because cross-entropy updating (which is characteristic of objective Bayesianism) also allows for explicit constraints which may be given by background knowledge. However, the general problem of representing ignorance and uncertainty is not solved either. Seidenfeld (1979b) gives an instructive example where Bayesian conditionalization is compared to cross-entropy updating in the presence of a nuisance parameter whose value is not known. More precisely, we are interested in the mean of a distribution μ, the variance being the unknown nuisance parameter. Starting with an (improper) uniform prior distribution, cross-entropy updating (see chapter 4) leads, after a series of i.i.d. trials, to a normal distribution for μ. By contrast, Bayesian conditionalizers who are, by construction of the example, supplied with more information, instead end up with Student's t-distribution as the posterior distribution. This is quite sensible as the t-distribution is closely related to the distribution of the mean μ when the variance is unknown. Moreover, by adding more information (e.g. the true value of the nuisance parameter), the t-distribution will be transformed into a normal distribution for μ, too. Those technicalities set aside, Seidenfeld's point can be stated thus: cross-entropy updating ranks "the normal and t-distributions in reverse order of informational content as they would be ranked by conditionalization."[5] The objective Bayesian neglects the interaction between two kinds of uncertainty: uncertainty about the parameter of interest and uncertainty about the nuisance parameter. The marginal distribution for the parameter of interest is actually not independent of the nuisance parameter, but cross-entropy updating treats it as if that were the case. Therefore, cross-

[5] Seidenfeld 1979b, 433.

entropy updating results in a "probability function that represents a state richer in empirical content than the belief state targeted for representation."[6] At least when nuisance parameters are involved, it seems to be difficult to represent uncertainty and ignorance by means of probability distributions, posing serious problems for objective Bayesians.

Despite these methodological concerns and the charge of subjectivity, Bayesian inference has proven to be incredibly fruitful in recent statistical research; and the intuitive idea behind it makes it easily understandable, applicable and compensates for the lack of complete objectivity. I would now like to pin down some foundational principles of statistical inference that are characteristic of, but more general than the Bayesian paradigm of statistical inference and that were developed by Allan Birnbaum (1962, 1972).[7] Birnbaum aims at a characterization of the evidential impact that an observation x generated by experiment E exerts on the parameter of interest ϑ. Such an 'experiment' corresponds, for instance, to a special experimental setup. Birnbaum wants to characterize (and to rule out) factors which should (not) affect our inference about ϑ. To this end, he introduces the function $Ev(E, x)$ as the "*evidential meaning*"[8] of the experiment. The essential properties of this Ev remain vague and to be clarified, but for the moment, it is sufficient to know that Ev summarizes all factors in (E, x) that are relevant for our inference about ϑ.

Birnbaum's first principle which is unanimously accepted by statisticians of all shades is the *Sufficiency Principle.* A *statistic* $T : \mathcal{X} \to S$ is any function from the sample space $\mathcal{X}$ to another measurable space S (e.g. the real numbers with their Borel sets). Such a statistic is called *sufficient* if the conditional distribution of the full data X, given the value of T, is independent of the parameter of interest ϑ, i.e.

$$P(X = x|T(X) = t, \vartheta) = P(X = x|T(X) = t). \tag{6.1}$$

In other words, given the value of T, the full data do not depend any more on ϑ. Thus, the parameter ϑ affects the data only in so far as it affects the sufficient statistics. By a simple application of Bayes's theorem, it can be

[6] Seidenfeld 1979b, 433.

[7] See Berger and Wolpert 1984 and Edwards 1992 for more recent versions and a discussion of those principles.

[8] Birnbaum 1962, 270.

shown that (6.1) is equivalent to

$$P(\vartheta|X = x) = P(\vartheta|T(X) = t). \tag{6.2}$$

Thus, if T is a sufficient statistic, the conditional distribution of ϑ is the same when conditioning on the full data X or on the transformation $T(X)$. In a subjective interpretation, our posterior degree of belief in ϑ and our inference about ϑ will not be changed by working with $T(X)$ instead of the full information X. This gives another rationale for basing one's inference on sufficient statistics since they will not distort the content of the data and suppress any important information. In total, working with sufficient statistics is literally spoken sufficient for inference from data:

> **Sufficiency Principle (SP):** "If E is a specified experiment, with outcomes x; it $t = T(x)$ is any sufficient statistic; and if E' is the experiment, derived from E in which any outcome x of E is represented only by the corresponding value $t = T(x)$ of the sufficient statistic; then for each x, $Ev(E, x) = Ev(E', t)$."[9]

Birnbaum's second principle is the Conditionality Principle which asserts the irrelevance of experiments not actually performed. I forego Birnbaum's technical formulation in favor of a more informal one:

> **Conditionality Principle (CP):** If the actually conducted experiment E is chosen from a collection of experiments $\mathcal{E}$ *in a way that is independent of the parameter* ϑ, then all other experiments can be neglected.[10]

In other words, our actual inference about ϑ does not depend on what we could have observed if other experiments had been conducted instead of the actual one. We should merely focus on what the actual observations tells us. The rationale for accepting conditionality is quite obvious – the evidence is a function of actual observations and should not depend on which experiments might have been performed instead. Therefore the CP is sometimes referred to as the 'Principle of Actuality' (see Berger and Wolpert 1984). Now, it is easy to see that both the Sufficiency and the Conditionality Principle are entailed by the following, stronger principle:

[9] Birnbaum 1962, 270.

[10] See Birnbaum 1962, 271.

> **Likelihood Principle (LP):** In an experiment E with observed data x, all experimental information *about* ϑ, is contained in the *likelihood function*[11] $\vartheta \mapsto P(x|\vartheta)$. All other information can be neglected. More precisely, if E and E' are two experiments and if the outcomes x and x' generate the same likelihood function, then $Ev(E, x) = Ev(E', x')$, without reference to the structure of E and E'.[12]

The Likelihood Principle entails that the probability of results *that could have been observed* is irrelevant to the statistical inference, as well as all other quantities that depend on the shape of the sample space $\mathcal{X}$. Notably, Bayesian inference conforms to the Likelihood Principle – it merely depends on prior probabilities (which have nothing to do with the experiment) and the likelihood of the data under the competing hypotheses, i.e. the likelihood function. But there are also statisticians who accept the Likelihood Principle without subscribing to Bayesian inference because they eschew the subjective components of Bayesianism. Those statisticians are *likelihoodists* (Hacking 1965, Royall 1997): contrastive relations of evidential favoring are more important than those of inductive support where alternative hypotheses need not always be specified. As already shown, favoring relations merely depend on the likelihood function of the observed data. The higher that likelihood, the more is a hypothesis favored over another one. As a measure of evidence, likelihoodists endorse the *likelihood ratio* between two competing hypotheses ϑ_0 and ϑ_1:

$$Ev_{\vartheta_1,\vartheta_0}(E, x) := \frac{P(x|\vartheta_1, E)}{P(x|\vartheta_0, E)}.$$

Likelihoodists thus have a tool for representing the evidential impact of observed data and may also be characterized as 'Bayesians without priors'. Despite this obvious advantage, the scope of likelihoodism is quite restricted. For instance, it is hard to express the likelihood ratio of composite hypotheses without introducing subjective priors – if one of the hypotheses is $H : \vartheta \in I$, I being a set of parameter values, then $P(x|H, E)$ cannot be computed without assigning a prior distribution to the single elements of H.

[11] In slight contrast to the philosophical terminology, the label 'likelihood function' is not a function of the data for fixed parameter: instead, it is a function on the parameter space that maps each ϑ to $P(x|\vartheta)$ where the observed data x are held fixed.

[12] See Birnbaum 1962, 271.

The Likelihood Principle is, for sure, quite strong, and may appear to be ill-motivated. However, it does not only entail both the SP and the CP – it is even *equivalent* to to their conjunction:

Theorem (Birnbaum 1962): *The Likelihood Principle is equivalent to the Conditionality Principle plus the Sufficiency Principle.*

Hence, those who are convinced by the reasons for accepting SP and CP have to swallow the LP, too. While many statisticians reject the Likelihood Principle, it is much more difficult to find statisticians who reject the Sufficiency or Conditionality Principle, putting deniers of the LP into a dilemma. Therein lies the significance of Birnbaum's theorem. Those who do not want to accept the LP usually choose to reject the CP. In the next section, I introduce a school of statistical inference which emphatically opposes the LP and the CP – error statistics.

6.3 Error statistics

In the early twentieth century, statistical theory rapidly developed – mainly, but not exclusively due to the enormous contributions of Ronald A. Fisher, Jerzy Neyman and Egon Pearson. The latter two established a theory of statistical testing known as Neyman-Pearson statistics. Error statistics is a school of statistical inference that blends Neyman and Pearson's statistical testing theory with ideas by Ronald A. Fisher (significance testing, rejection trials). Therefore, I first explain the basic principles of Neyman-Pearson statistics.

The main idea of Neyman-Pearson statistics can be illustrated when two mutually exclusive statistical hypotheses H_0 and H_1 are tested against each other. At the end of the test, we are supposed to make a decision in favor of one of the hypotheses, i.e. we either accept H_0 and reject H_1 or we accept H_1 and reject H_0. Naturally, the decision procedure is supposed to be *reliable*, i.e. it should guide us towards the true hypothesis in the vast majority of cases. (Assume for reasons of simplicity that either H_0 or H_1 is true.) This is a position similar to epistemic externalism – beliefs are justified if they are generated by reliable processes, i.e. processes that tend to generate much more true than false beliefs. Here *error probabilities* come into play. In the error-statistical framework, statistical tests (=decision procedures) are characterized by their probabilities to accept H_1 if H_0 is true and their probabilities to accept H_0 if H_1 is true. These are the two characteristic tendencies to make a wrong decision. How does this work in practice? Consider a Bernoulli trial with merely two possible outcomes: success and failure. Examples are the repeated toss of a coin and similar experiments where only two outcomes are possible. A sequence of independent and identically distributed Bernoulli trials with a fixed sample size n is called a *Binomial experiment* because the number of successes k is distributed according to the Binomial distribution with density

$$B_{n,p}(k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k \, (1-p)^{n-k}.$$

This means that if the tendency to get a success in a single trial is p, the probability to get k successes after n trials will be equal to $B_{n,p}(k)$. Of course, the $B_{n,p}(k)$ sum up to 1 in total.

Now consider the following situation.[13] Someone (let's call him S) claims to possess the ability of extrasensory perception (ESP). As a proof of his ability, S claims to be able to find out a particular playing card (say, the ace of spades) out of two reversed cards with a success frequency that is significantly higher than pure chance. To find out whether S's claims are correct, we present him two cards upside down, one of which is the ace of spades and let him guess. Then we note success or failure. This trial is repeated twenty times and the number of successes follows the Binomial distribution. Now, we have two hypotheses about S's success probability in such an experiment. On the one hand, we have the *default* or *null hypothesis* (briefly: null) or H_0 that S is merely guessing and not better than ordinary human beings: $H_0 : p = 0.5$. On the other hand, we have the alternative hypothesis that S has ESP and that he has a significantly higher success probability than pure chance would suggest: $H_1 : p = 0.7$. A comparably high success frequency cannot be explained by normal means if the experiment is properly conducted. Now, we have to choose a decision rule. Due to the potential loss of reputation, we have to avoid an unjustified ascription of extrasensory perception more than we have to avoid an incorrect decision for the default hypothesis that S has no extrasensory abilities. This illuminates the point of the label 'null/default hypothesis'. In particular, we would like to make an erroneous decision in less than 5% of all cases when H_0 is true. Hence, we decide to opt for H_1 if

[13] The example is taken from Georgii 2002, 248-49.

and only if S is right in at least 15 of 20 trials. See figure 6.1 for an intuitive justification of that rule: the higher the probability of an observation under H_1 as compared to H_0, the more we are willing to reject H_0.

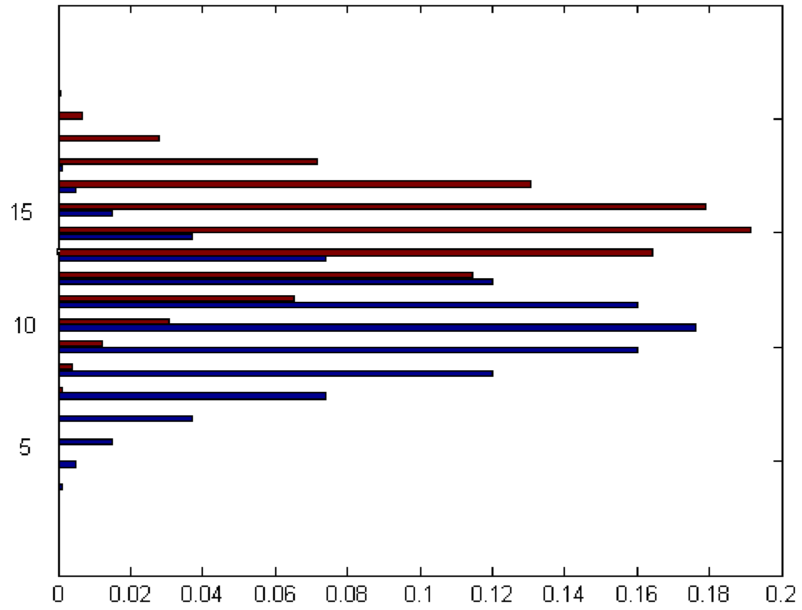

Figure 6.1: The probability densities of the Binomial distribution for H_0 : $B_{20,0.5}$ (dark bars) and H_1 : $B_{20,0.7}$ (light bars).

Having formed a decision rule, we can calculate the probability of an *error of the first kind* – to opt for H_1 although H_0 is true. It is the probability to observe 15 or more successes when H_0 is right:

$$\begin{aligned} P(\text{decision for } H_1|H_0) &= P_{H_0}(k \geq 15) \\ &= \sum_{k=15}^{20} \left(\begin{array}{c} 20 \\ k \end{array} \right) 0.5^k \, (1-0.5)^{20-k} \\ &= 0.0207. \end{aligned}$$

Hence the error the first kind, the probability of an erroneous decision for H_1 is indeed very small – it is just above 2 percent. On the other hand, the *error of the second kind* – the probability of an erroneous decision for H_0 is equal to

$$\begin{aligned} P(\text{decision for } H_0|H_1) &= P_{H_1}(k \leq 14) \\ &= \sum_{k=1}^{14} \left(\begin{array}{c} 20 \\ k \end{array} \right) 0.7^k \, (1 - 0.7)^{20-k} \\ &= 0.5836. \end{aligned}$$

This is clearly above 50%, indicating that in more than half of the cases, the alternative hypothesis will not be detected even if it is true. This sounds very high, but we have to recall that a decision for H_1 might put to jeopardy a lot of the experimenter's reputation, given the widespread sceptical doubts towards extrasensory perception. In the present example, it is thus more important to control the error of the first kind. The name 'error statistics' is indeed derived from the desire to control the error probabilities in choosing a statistical decision rule. The lower the error probabilities, the more reliable the statistical test. Note that error probabilities are not subjectively interpreted probabilities (i.e. some kind of personal credences), but *objective* probabilities – probabilities of observing a given (set of) event(s) under a specific hypotheses. Thus, the recourse to rational degrees of belief that was prevalent in the Bayesian approach is avoided.

The problem with Neyman-Pearson statistics is that it is a theory of statistical decision rules – it compares the reliability of various testing procedures between which we can select. But Neyman-Pearson theory does not give (at least, not at first sight), a post-experimental quantification and representation of the observed evidence. Scientists are often more keen on measures of evidence than on the properties of specific decision rules that neither quantify the strength of evidence against the rejected hypothesis nor give a post-observational assessment of the tenability of the rejected hypothesis. For this reason, Deborah Mayo and Aris Spanos, two leading proponents of the error-statistical approach, have suggested that not only the *result* of a hypothesis test be reported, but also the *degree of severity* with which the accepted hypothesis has passed the test (Mayo 1996, Mayo and Spanos 2006). This degree of severity is supposed to quantify the observed evidence:

Definition 6.1 *"A statistical hypothesis H passes a* ***severe test*** *T with data x_0 if,*

- *x_0 agrees with H, and*

- *with very high probability, test T would have produced a result that accords less well with H than x_0 does, if H were false.*"[14]

Clearly, the dependence on 'counterfactual experiments' expressed in the second clause implies the violation of the Conditionality Principle. Thus, the error-statistical approach stands in sharp contrast to the Likelihood Principle and Bayesian inference. Moreover, the two clauses in the definition of a severe test rely both on a measure of concordance and dissent between the observed data and the two competing hypotheses. This requires the definition of a statistic that measures in how far the observed result x_0 diverges form H (in the direction of $\neg H$). Furthermore, it is clear that the actually observed result x_0 plays a crucial role in determining whether H has passed a severe test. Similar to Neyman-Pearson tests, error statistics justifies endorsal of a hypothesis H by the reliability of the conclusion-generating procedure: it would have been very unlikely to obtain a result fitting so well (namely, x_0) with hypothesis H if H had been wrong.

The standard statistic for ordering the sample space elements according to their discrepancy to H is the likelihood ratio between $\neg H$ and H or a monotonous function thereof. The higher this value, the more do the observed results diverge from H and the closer they are to $\neg H$. In the ESP example, the number of successes of the person would be the most natural choice. The more successes the person scores, the closer are the results to the alternative and the more distant are they from the null. Now, the second clause of definition 6.1 suggests that the degree of severity with which a hypothesis passes a test be measured by the probability to observe a more diverging result in case the hypothesis is false. Indeed, this is the definition suggested by Mayo and Spanos in their 2006: The severity $s : \Theta \times X \rightarrow [0, 1]$, X being the sample space and Θ being the hypothesis space, is defined as

$$s(H, x_0) := P(d_H(X) > d_H(x_0) \mid \neg H). \tag{6.3}$$

where d_H measures the distance between the observed result and H. In the ESP example, after $x_0 = 14$ successes and six failures, the default hypothesis H_0 (the person has no ESP) is accepted by the testing procedure, but it passes the statistical test with quite low severity:

$$\begin{aligned} s(H_0, x_0) &= P(X > x_0 \mid H_1) = P(X > 14 \mid p = 0.7) \\ &\approx 0.416. \end{aligned} \tag{6.4}$$

[14] Mayo and Spanos (2006, 329), italics in the original.

The low severity indicates that the result is quite weak evidence and does not remove the uncertainty about the person's abilities, quite in agreement with our intuitions that 14 successes are quite impressive. But if we had observed $x_1 = 15$ successes, we would have rejected H_0 and H_1 would have passed the test with high severity:

$$\begin{aligned} s(H_1, x_1) &= P(X < x_1 \mid H_0) = P(X < 15 \mid p = 0.5) \\ &\approx 0.979. \end{aligned} \tag{6.5}$$

Thus, if 15 success were observed, we could confidently assert H_1. By taking into account post-experimental quantifications of evidence that accompany the result of statistical tests, the error-statistical approach blends and refines both the Neyman-Pearson and the Fisherian approach to statistical inference. Neyman-Pearson testing suffered under the lack of post-observational measures of evidence whereas Fisher's approach to statistical inference is objectionable because alternative hypotheses are not explicitly taken into account (more on this in the next chapter). Error statistics combines both approaches into a unified scheme of statistical inference without subjective probabilities and in sharp contrast to the Likelihood Principle. We will now see how the Neyman-Pearsonian emphasis on the predesignation of statistical tests extends to error statistics and how it helps to distinguish spurious from relevant correlations.

6.4 Error statistics and predesignation

Neyman-Pearson theory is a predesignationist theory of statistical inference, in the sense that the decisions must be executed and the error probabilities must be reported in the way that the procedure was designed. To see the point of this remark, let us come back to our example about the person with alleged extrasensory perception. There is a tempting kind of reasoning which is not compatible with the error-statistical principles. Assume that in the ESP example, we have observed 14 successes and only 6 failures. Certainly, this is a remarkable success rate which casts some doubt on the null hypothesis that S is just guessing. Moreover, the severity with which the No-ESP-hypothesis passes the test is quite low. Strongly impressed by S's performance, the director of the experiment decides to modify her decision rule *post experimentum* as to accept H_1 (the alternative) for 14 or more successes, instead of demanding 15 or more successes. Consequently, she actually accepts H_1. Then, the error of the first kind is 0.0577 which is still quite low and the error of the second kind is 0.392 which seems to be a substantial improvement. But here, we see that Neyman-Pearson statistics is a pre-experimental framework. The director's reasoning points out that she factually adopted another cutoff point for rejecting H_0 than she claimed at the outset. If S scores merely 14 successes, she accepts the hypothesis that he has ESPs and reports an error probability of 0.0577 (and a degree of severity of 0.9423). But if S had scored 15 or more successes she would also have ascribed extrasensory perception to S, and have reported the 'old' error probability of $\alpha \approx 0.02$, as if she had accepted the null for $k = 14$. Her sudden change of mind in the case of $k = 14$ biases the experimental report in favor of H_1 because the error probabilities for $k \geq 15$ were incorrectly reported. Neyman-Pearson theory is a theory of predesignated statistical decision rules, and one must not *post mortem* modify those rules and report rules which differ from those one had originally in mind.

There is a more relevant and maybe even more perspicuous example for the vices of violating the predesignationist stance of Neyman-Pearson statistics. Assume twenty variables $X_1, \ldots, X_{20}$ are tested for correlation with a response variable Y. We take a number of samples from each variable and a computer program checks the data for significant correlations. We find out that only the variable X_1 is strongly correlated with Y – actually, the hypothesis that X_1 and Y are independent would be rejected at the 0.05 level by a statistical test. In other words, a test with a (first kind) error probability of 0.05 would lead to a rejection of the independence hypothesis. Now, it is tempting to claim that a strong correlation between X_1 and Y has been found and has been verified at the error level of 0.05. Furthermore, we would report the observed degree of severity. But in fact, we would run into a dangerous pitfall since we did not perform a proper test but we merely scanned the data for correlations.

> "When hypotheses are tested on the same data that suggested them and when tests of significance are based on such data, then a spurious impression of validity may result. The computed level of [severity, J.S.] may have almost no relation to the true level."[15]

[15] Selvin 1970, 104. I have replaced 'level of significance' by 'level of severity' in order to keep a consistent terminology.

How can we make this point precise? In checking twenty different input variables for correlations with the response variable Y, we will sooner or later discover a correlation for one of the input variables *even if there is no real correlation.* This is just a sampling effect – if you sample on long enough (here, 20 times), at least one of a set of individually unlikely events will eventually happen with high probability. Hence, in scanning the data we did not actually test the hypothesis that X_1 and Y are correlated but the hypothesis that at least one X_j, $j \in \{1, \ldots, 20\}$, is correlated with Y since we would have reported a correlation regardless of the value of j. This makes quite a difference: if we select X_1 beforehand (i.e. before peeking at the data) and test it for correlation, the error probability is indeed 0.95 = 1-0.05, but if we decide to scan the data for correlations and report those which are 'significant at the 0.05 level', we end up with a poor error probability: The probability of observing a strong correlation on one of the 20 factors given that there is no genuine correlation is 0.64. Thus, in reporting error probabilities and degrees of severity we have to correctly identify the procedure which was used:

> "if you change the test procedure the error probabilities change, and if you report significance levels in the usual way [...] then you are going to get your error probabilities wrong."[16]

Here we see another time that Neyman-Pearson testing theory relies on predesignation and faithful adherence to the decision procedure. Violating the predesignationist stance makes us overly optimistic: we think spurious correlations are genuine and claim that hypothesis can be rejected at high significance levels/low error probabilities when they can't. Thus, we bias our error probabilities and lead the entire inference astray:

> "since violating predesignation may alter the actual significance level (by altering the test procedure), it is invalid to report the results *in the same way* as if hypotheses were predesignated.[...] If one fails to heed [the warning], tests will be construed erroneously as having high severity."[17]

Neyman-Pearson theory is thus able to explain what is going wrong when scientists hunt a huge bulk of data for significant results and publish them without mentioning the hunting procedure: they do not pay attention to chance effects which will inevitably occur in a large set of data.

[16] Mayo 1996, 311.

[17] Mayo 1996, 317, original emphasis.

6.5 Summary

This chapter has motivated why a special focus on statistical hypotheses is important in examining theories of inductive inference. In particular, the rapidly progressing introduction of probabilistic models in the empirical sciences asks for a clarification of foundational methodological issues. Statistical techniques are present in the formulation and analysis of models, in estimating unknown parameters, testing hypotheses and simulating real-world processes, in other words, in all aspects of the interpretation of data. The chapter has contrasted the two major schools to statistical inference, the Bayesian and the error-statistical approach. The focal point of their disagreement was the Likelihood Principle.

Bayesian statistics is a natural part of the Bayesian theory of inductive inference which we have encountered in the previous chapters. Still, it is a partially subjective theory, and some scientists might feel uncomfortable with abandoning the pretensions for absolute objectivity and working with rational degrees of belief instead. The various attempts to objectify Bayesian inference by means of 'objective prior distributions' are often practically useful, but representing lack of information and ignorance by special probability distributions blurs the distinction between risk and uncertainty. Both the personalist and the objectivist version of Bayesianism have their virtues and vices.

The Likelihood Principle (LP) is a principle of statistical inference which asserts that the evidential content of the data is captured in the likelihood function. Birnbaum (1962) has shown that the Likelihood Principle is equivalent to the conjunction of two more fundamental and intuitive principles, making it hard to reject the LP. It forms a part of Bayesian and likelihoodist statistical inference whereas statisticians in the Neyman-Pearson-Fisher tradition (e.g. error statisticians) deny it.

Error statistics is the label for a theory of statistical inference that mainly builds on Neyman and Pearson's theory of testing alternative hypotheses and connects this theory with post-experimental quantifications of the observed evidence. Here, *degrees of severity* (Mayo and Spanos 2006) play a crucial role. Like Neyman-Pearson theory, error statistics is a predesignationist theory of inference, in the sense that it is mandatory to stick to predesignated decision rules in order not to bias the final conclusions. This last point is illustrated in a series of examples. In the next chapter, we will devote more space to the controversy between the error-statistical and the Bayesian approach: we will compare the different conceptions of evidence as well as dissolve the dissent about the relevance of experimental design.

Chapter 7

Evidence and Design

Which relevance does the *design* of a statistical experiment in science have, once the experiment has been performed and the data have been observed? Do data speak for themselves or do they have to be assessed in conjunction with the design that was used to generate them? Few questions in the philosophy of statistics are a subject of greater controversy. The debate about the inferential role of experimental design standardly narrows down to the inferential role of *stopping rules* that describe under which circumstances an experiment has to be terminated. If these rules were relevant to the interpretation of an experiment, it would be mandatory to fix them in advance, i.e. before actually conducting the experiment. That would have severe implications for scientific practice and affect the way data are collected and experimental reports are written. Hence, both scientists and philosophers of science should pay high attention to the role of stopping rules in statistical inference.

The classical example for an application of stopping rules are *sequential trials*. Sequential trials repeat a single experiment, accumulating evidence from several independent and identically distributed trials. They can be compared to the repeated toss of a coin and are standardly applied when testing medical drugs and giving them to a group of patients. Possible stopping rules could then be 'give the drug to twelve patients', 'give the drug until the number of failures exceeds the number of recoveries' or 'give the drug until funds are exhausted'. In this example, we recast the question about the relevance of experimental design as the question whether the *evidence* about the effectiveness of the drug is sensitive to the proposed ways to conduct the experiment.[1] In other words, we would like to investigate whether stopping rules are evidentially and inferentially relevant.

This topic of inquiry is closely related to the debate about measures of evidence. Obviously, asking for the evidential relevance of experimental design implies that the answer depends on what we expect from a measure of evidence and which measures of evidence should be preferred. Measures of evidence provide a way to transform the observed data into a basis of meaningful scientific inference. Thus, they play a crucial role in post-experimental analysis. Here, I would like to take a twofold perspective and to review the problem of finding an adequate measure of evidence from a foundational as well as a practical point of view, i.e. adequate measures of evidence should be applicable and fruitful tools for scientists who work with statistical methods. These results will then backfire on the evidential relevance of experimental design. The two conflicting positions in the debate are associated with the error-statistical and the Bayesian school of statistical inference. We will present both view in details and then try to adjudicate between them.

7.1 The controversy about experimental design

Neyman-Pearson statistics is concerned with statistical testing and the comparison of two mutually exclusive hypotheses (called the *null hypothesis* H_0 and the *alternative* H_1). After looking at the data, one of them is accepted and the other one is rejected. The data are thus used to decide between the two hypotheses. Tests are ranked according to their error probabilities, i.e. the probability of erroneously opting for the alternative and the probability of erroneously opting for the null hypothesis.[2] We describe such tests by the tuple $\langle \alpha, \beta \rangle$ which encodes the probability of erroneously rejecting the null (α) and the probability of erroneously accepting the null (β). Error statisticians say that the lower the error probabilities, the higher the severity

[1] Technically, stopping rules are integer-valued random variables τ, i.e. functions from the sample space into the set of natural numbers. They indicate the number of repetitions of the trial as a function of the observed results. We strictly confine ourselves to *noninformative stopping rules* – stopping rules that are independent of the prior distribution of the parameter.

[2] Other Neyman-Pearson procedures (parameter estimation, construction of confidence intervals) are equally justified by the error probabilities which characterize that procedure.

with which the accepted hypothesis has passed a test. In other words, error probabilities give a benchmark for the severity and reliability of statistical inference.

Now, it is interesting to note that error probabilities depend on results that *could have been observed under the actual experimental design.* For instance, imagine that we would like to find out the probability of recovery of when giving a drug against a specific disease. If the patients are not treated, only half of them recover. We make some simplifying assumptions and set up the experiment as a Bernoulli trial, i.e. for any patient there are only two possible results (recovery/no recovery). Each patient has the same probability of recovery ϑ when (s)he is given the drug, and the results for each patient are independent of each other – in other words, we have a sequence of i.i.d. Bernoulli trials. As our default or null hypothesis we choose the claim that the new drug is no more effective than a placebo, i.e. that the probability of recovery is equal to 1/2 ($H_0 : \vartheta = 0.5$). The alternative hypothesis posits that the new treatment is more effective than a placebo ($H_1 : \vartheta > 0.5$).[3] Now compare the following two stopping rules:

- τ_1: Give the drug to exactly twelve patients (fixed sample size).
- τ_2: Give the drug until three patients have failed to recover.

τ_1 invokes an n-fold *Binomial* experiment, τ_2 invokes a *negative Binomial* experiment. Both designs are somewhat plausible: the Binomial design bounds the total number of trial persons, the negative Binomial design bounds the number of failures in the treatment group. We would like to test the hypothesis H_1 with severity $\alpha = 0.05$, i.e. only in 5% of all cases where H_1 is false (and H_0 is true) we erroneously reject H_0 and opt for H_1. According to the Neyman-Pearson Lemma, there are uniformly optimal tests for testing H_0 against H_1, i.e. tests that minimize β for fixed α, regardless of the value of ϑ. In the remainder, we assume that these tests are adopted in both designs. Now, the tests are started, but (as interested readers of a medical research report) we do not know the experimental design – in particular, we do not the stopping rule – and are told that 12 patients have been examined and that three patients did not recover (among them the 12th patient). Now, the Binomial design and the negative Binomial design yield different decisions.

[3] Assume that it can be ruled out that the drug is conducive to the disease ($\vartheta < 0.5$).

Let k denote the number of recoveries and n the number of trials. In the Binomial design, due to

$$B_{12,0.5}(\{k \geq 9\}) = \frac{299}{4096} \approx 0.073$$

the actually observed result $k = 9$ lies in the *acceptance area*, i.e. H_0 is accepted whereas in the negative Binomial design,

$$\overline{B_{3,0.5}}(\{n \geq 12\}) = \frac{67}{2048} \approx 0.033$$

so that $n = 12$ lies in the *rejection area* and H_0 is rejected in favor of H_1. In a Neyman-Pearson inference, different experimental designs curve the sample space differently so that the associated optimal tests yield different decisions for the same actual result. Hence, for the reader of the medical research report, it is absolutely essential to know which design was adopted in order to understand the scientists' decision whether to accept or to reject the null hypothesis. While this may sound awkward at first sight, there is also a rationale for this property of hypothesis tests: some designs (as the negative Binomial design) are slightly biased towards a particular hypothesis since the last observed patient in the row has to be a non-recovery. Producing an experimental result that is favorable to H_1 is therefore harder in a negative Binomial design than in a Binomial design. A more everyday example may illustrate that point: If a football match were terminated at once if a certain team took the lead, the opponents would complain that the design of the match was unfair to them and that the point of termination should not depend on the current score. Since experimental design and stopping rules affect error probabilities and the outcome of hypothesis tests, Neyman-Pearson and error statisticians conclude that experimental design is inferentially relevant and crucially affects statistical inference.

By contrast, Bayesians and likelihoodists deny the inferential relevance of experimental design. This is a direct consequence of the Likelihood Principle (LP), which both statistical schools accept and which rests on the more primitive Sufficiency Principle and Conditionality Principle (see the previous chapter). In particular, all inference about the unknown parameter merely depends on the likelihood function of the observed result: If ϑ is our parameter of interest and x is the observed result, our inference depends merely on $P(x|\vartheta)_{\vartheta \in \Theta}$ and not on other results that could have been observed. Indeed, both designs impose the same likelihood function (up to a constant factor).

This has direct consequences for the role of stopping rules and leaves no room for the inferential relevance of experimental design. That view can be condensed into the following principle:

> **Stopping Rule Principle (SRP):** In a sequential experiment with observed data $x^{(n)} = (x_1, \dots, x_n)$, all experimental information *about* ϑ is contained in the function $P_n(x^{(n)}|\vartheta)$; the stopping rule τ that was used *provides no additional information about* ϑ.[4]

But how does the SRP deal with the error-statistical charge that the fairness of experimental design is important to the interpretation of an experiment? Adherents of the SRP and the LP do not feel responsible for that problem. For them, data have an evidential content which is not affected by the choice of a stopping rule which is a 'mere intention' in the head of the experimenter. For example, if the experimenters in the above medical trial had forgotten to fix a stopping rule, the experimental results would be uninterpretable for error statisticians. For defenders of the LP and the SRP, this is unacceptable and a serious drawback of Neyman-Pearson ('classical') statistics:

> "The irrelevance of stopping rules is one respect in which Bayesian [and likelihoodist, J.S.] procedures are more objective than classical ones. Classical procedures [...] insist that the intentions of the experimenter are crucial to the interpretation of the data."[5]

Indeed, in a Bayesian measure of evidence and more generally, in a Bayesian inference, stopping rules do not play a role. As a measure of evidence, Bayesians standardly use Bayes factors, the ratio between prior and posterior odds which generalizes the likelihood ratio:

$$B(H_1, H_0, x) := \frac{P(H_1|x)}{P(H_1)} \frac{P(H_0|x)}{P(H_0)} = \frac{\int_{H_1} P(\vartheta)\, P(x|\vartheta)\, d\vartheta}{\int_{H_0} P(\vartheta)\, P(x|\vartheta)\, d\vartheta}. \tag{7.1}$$

Eventually, the entire Bayesian inference only depends on the prior probabilities of the hypotheses and the likelihood of the data under the competing hypotheses. Both quantities are independent of the stopping rule.

To settle the dispute between Bayesians and likelihoodists on the one side and Neyman-Pearson and error statisticians on the other side is no easy task,

[4] Berger and Berry 1988, 34, italics in original, notation changed for convenience. The first formulation of the SRP goes back to Barnard, see e.g. his 1949.

[5] Edwards et al. 1963, 239.

however. Typically, statisticians and philosophers of science in the Neyman-Pearson tradition accuse the Bayesians of wearing Bayesian glasses and being unable to see the problems associated with neglecting stopping rules in the interpretation of an experiment.[6] Indeed, Bayesians often argue for the Stopping Rule Principle on foundational grounds: rejecting the SRP implies the violation of either the Sufficiency Principle or the Conditionality Principle which are both very plausible.[7] Here, basic intuitions are invoked and it is hard to convince those who do not share such intuitions. Vice versa, the arguments against the SRP are based on an error-statistical understanding of evidence and statistical inference. Bayesians generally deny the validity of error-statistical arguments.[8] Since both sides tend to presuppose what is at stake, the debate seems to be in a stalemate.

Of course, it is not possible to re-invent a debate that has been ongoing for several decades. But I think that the existent arguments could be structured in a better way. Here is my project: First, it is asked which criteria a measure of evidence *suitable for scientific use* should satisfy. Thus we combine arguments from mathematical statistics with a methodological perspective on the needs of experimental practice. Second, we check whether error-statistical measures of evidence satisfy those criteria and how they perform in other respects. This will eventually establish the inadequacy of such measures and elucidate the comparative character of evidence measures. Third, we try to readjust the function of putative error-statistical measures of evidence in statistical inference. Fourth, we infer from the previous results to the evidential irrelevance of evidential design. This conclusion is integrated into a decision-theoretic perspective and defended against classical counterarguments. Fifth and last, we stress that it would be unwise to assert that careful experimental design is negligible in science: Scientific inference arguably builds on more factors than statistical evidence. When the cost of a single observation in a sequential trial is substantial, experimental design helps to control the costs of the experiment. Careful design is used to optimize the tradeoff between scientific insights and experimental costs and therefore indispensable for conducting sequential trials.

[6] See Mayo 1996, 348, and Mayo and Kruse 2001.

[7] See Berger 1985, 507-509.

[8] See Berger and Berry 1988, 45 and in a similar vein, though from a likelihoodist point of view, Royall 1997, 68-71.

7.2 Measures of evidence and p-values

Evidence about a parameter is required for inferences about that parameter, e.g. for sensible estimates and decisions to work with this rather than that value. An evidence measure transforms the data as to provide the basis for a scientific inference. In order to be suitable for public communication in the scientific community and use in research reports, a measure of evidence should be free of subjective bias and distortion. While we might disagree on the a priori plausibility of a hypothesis, we should agree on the strength of the observed evidence. Scientists and policy-makers often want to make evidence-based decisions, but to take an evidence-based approach to statistical inference seriously presupposes consent on what the data tell us.[9] Therefore we need a method to quantify the information which the data convey that is independent of idiosyncratic convictions and immune to deliberate manipulations. We will now see whether error-statistical measures are able to fulfil that task.

Error statisticians try to flesh out a full theory of statistical inference on the basis of Neyman's and Pearson's work on statistical testing: Two mutually exclusive hypotheses (the null hypothesis and the alternative) are compared to each other, and after looking at the data, one of them is accepted and the other one is rejected. The data are thus used to decide between the two hypotheses. Such tests are ranked according to their reliability, i.e. their ability to guide us towards the true hypothesis. The benchmark for the reliability of a Neyman-Pearson test are the *error probabilities* – the probability of erroneously rejecting the null in favor of the alternative (α) and the probability of erroneously accepting the null hypothesis (β). Neyman-Pearson theory is essentially a theory of *statistical decisions*, so it is quite natural to report the properties of the decision procedure. While the error probabilities give useful pre-experimental information about the reliability of a statistical test which is going to be used, their post-experimental interpretation is much more difficult: First, two results in the rejection area can have different 'discrepancies' to the null hypothesis, e.g. a result close to the acceptance/rejection cutoff seems to be weaker evidence against H_0 than a result far in the rejection area. Nevertheless, for the error probabilities of a predesignated test, this difference is irrelevant. Second, there are severe con-

[9] To a certain extent, this aspect of the word 'evidence' is contained in the similarity to the word 'evident'.

ceptual problems regarding the post-experimental use of error probabilities: If H_0 was rejected and the test had a probability of an erroneous rejection equal to $\alpha = 0.05$, this does *not* imply that we made the right decision with a probability of 0.05. The null hypothesis is either true or false, but not with a probability of 0.05 – error-statisticians sneeze at using probabilities in the subjective sense. Hence, the evidential, post-observational meaning of error probabilities is not clear. Confidence interval construction gives a paradigmatic example: A confidence interval $[x, y]$ for a parameter of interest ϑ either contains or does not contain the true parameter value. Again, for an error statistician, it is not possible to assign any degree of confidence to the statement that $\vartheta \in [x, y]$. So what is the meaning of the probability α used to qualify confidence intervals? In fact, this α only describes properties of the *construction procedure*, but it does not state whether we are *factually* justified in believing that $\vartheta \in [x, y]$ (for this, we would need the prior distribution of ϑ). But it is precisely the question of factual justification that experimenters are interested in, as witnessed by Pratt's description of the dilemma:

> "We can say to an experimenter: 'A method yielding true statements with probability .95, when applied to your statement yields the statement that your treatment effect is between 17 and 29, but no conclusion is possible about how probable it is that your treatment effect is between 17 and 29'. The experimenter, who is interested not in the method, but in the treatment and this particular confidence interval, would get cold comfort if he believed it."[10]

Thus, it is not clear how the reliability of the statistical method transfers to the actual confidence which the experimenter should have in his interval. If confidence intervals are understood as matters of decision or statistical inference, error probabilities alone give a rationale for using them. Neither can error probabilities be used as a post-experimental, evidential basis for scientific inference. Neyman-Pearson statistics is a theory of statistical decision rules, but "it does not address the problem of representing statistical evidence"[11].

[10] Pratt 1961, 165.

[11] Royall 1997, 58

Scientists often ask for a post-experimental quantification of the evidential content of the data. Since Neyman-Pearson testing theory is unable to deliver such a quantification, it is in practice often blended with elements from Fisherian statistics and significance testing. The error-statistical paradigm accounts for that demands by means of degrees of severity. As mentioned in the previous chapter, error statisticians build on elements of the Neyman-Pearson theory of statistical testing and supplement them by inferential interpretations of statistical tests and post-data assessments of strength of evidence, i.e. the degree of severity with which a hypothesis passes a test. In practice, this function is most often taken by p-values which are closely related to degrees of severity. The p-value sums up the H_0-likelihoods of those observations that fit the null model to a lower degree than the observed value x_{obs}:

$$p_{\text{obs}}(x_0) := P(d_{H_0} \geq d_{H_0}(x_0) \mid H_0) \tag{7.2}$$

where d_{H_0} is again a statistic measuring the discrepancy between H_0 and the data, as known from the definition of degrees of severity.[12] In other words, p-values measure the probability that, if the null hypothesis were true, a more extreme result than the actual one would be observed. Often, they are also called the 'observed level of (statistical) significance'. Notably, the severity of experimental tests is closely related to p-values: In the above example, the severity with which $\vartheta > 0.5$ passes a severe test against $\vartheta = 0.5$ is $1 - p_{\text{obs}} = \mathbb{P}(\overline{X} < \overline{x_{\text{obs}}} | p = 0.5)$. In a little bit more detail:

$$\begin{aligned} s(\vartheta > 0.5,\, x_0) &:= P(d_{H_1}(X) > d_{H_1}(x_0) \mid \neg H_1) \\ &= P(d_{H_0} < d_{H_0}(x_0) \mid H_0) \\ &= 1 - p_{\text{obs}}(x_0). \end{aligned}$$

Thus, in our discussion of p-values, we also cover Mayo's degrees of severity. In statistical practice, p-values are often used to assess the tenability of a 'null' (or default) hypothesis H_0 in the light of observed data, low p-values speaking against the null. Indeed, they are widespread in the empirical sciences and often used as a summary of the evidential import of an experiment. For instance, the level of the p-value often decides whether or not an experimental result can be published, e.g. many psychologists consider only results with a p-value lower than 0.05 to be (statistically) significant and therefore

[12] Such a statistic $d_{H_0}(X)$ has to be minimal sufficient, i.e. representable as a function of any other sufficient statistic. More explanation follows in the text.

publishable. The precise role of p-values, however, is not clear, despite their enormous popularity. They are often confounded with posterior probabilities of a null hypothesis, e.g. when a p-value of 0.04 obtains, practitioners without a sufficient mathematical education often tend to assert that 'the null hypothesis has a probability of 0.04'. Although this is a well-known fallacy – p-values do not give posterior probabilities – practitioners often commit it. These pitfalls put aside, p-values are often cited as a basis for the rejection of a null hypothesis or, vice versa, for claiming that the evidence against the null hypothesis is not sufficiently strong to warrant assertion of the alternative. There is a telling relationship between p-values and error probabilities: a p-value lower than α means that the null hypothesis is rejected in a test with error probability (of the first kind) equal to α. Figure 7.1 illustrates this idea for testing $H_0 = N(0,1)$ against an unspecified alternative where d_{H_0} is identified with the probability density of H_0. By contrast, figure 7.2 exemplifies the one-sided testing problem where the null is tested against a specific alternative $N(1,1)$ and d_{H_0} is identified with the likelihood ratio between H_1 and H_0.

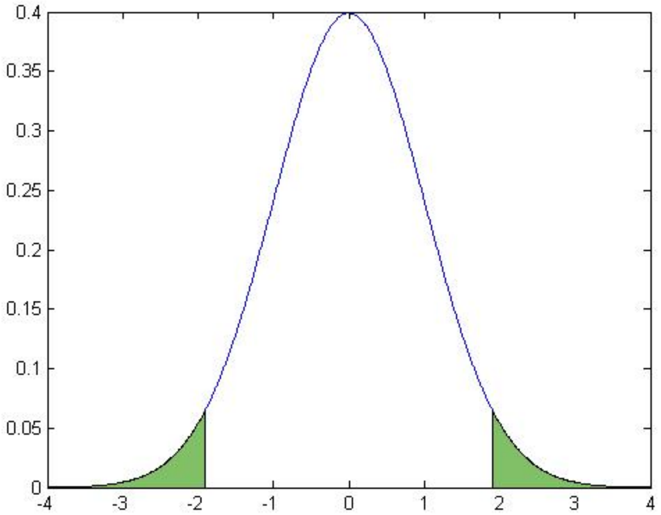

Figure 7.1: The null hypothesis $H_0 : N(0,1)$ (full line) is tested against an unspecified alternative. The shaded area represents the set of results where H_0 is rejected in a test with predesignated error probability of 0.05 and where an observation yields a p-value below 0.05.

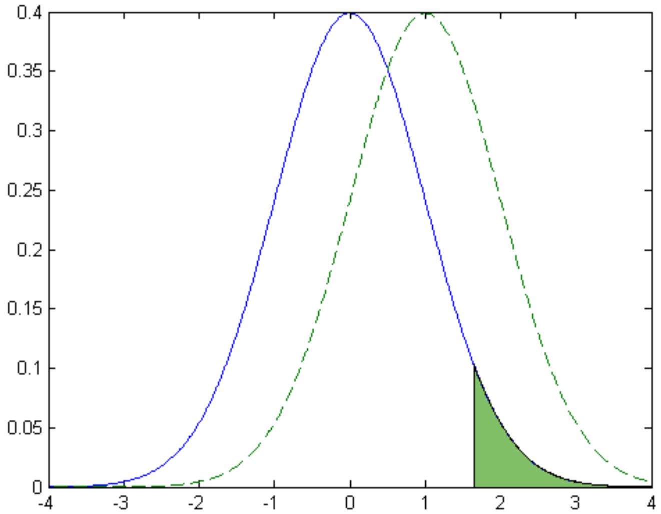

Figure 7.2: The null hypothesis $H_0 : N(0, 1)$ (full line) is tested against the alternative $H_0 : N(1, 1)$ (dashed line). The shaded area represents the set of results where H_0 is rejected in favor of H_1 in a test with predesignated error probability of 0.05 and where an observation yields a p-value below 0.05.

It is now suggestive to conclude that a low p-value suggests poor evidence for H_0 and that a high p-value suggests good evidence for H_0, independent of the specific alternative hypotheses. But this is a misunderstanding: a low p-value may indicate evidence against the null hypothesis, but a high p-value is not very telling. For instance, the observation $x_0 = 0$ leads to a maximal p-value ($p_{\text{obs}} = 1$), but while it *fits* the null well, it is obviously premature to say it is strong *evidence* for the null.

> "Although a significant departure [from the null] provides some degree of evidence against a null hypothesis, it is important to realize that a 'nonsignificant' departure does not provide positive evidence in favor of that hypothesis. The situation is rather that we have failed to find strong evidence against the null hypothesis."[13]

Due to the ubiquitous occurrence of p-values in scientific research reports, their inferential role is of keen and abiding interest. Here it is essential to note that p-values are sensitive to experimental design and stopping rules.

[13] Armitage and Berry 1987, 96.

This point was made very often in the existing debate (e.g. Howson and Urbach 1993, Royall 1997): p-values do not only depend on the probability of the actually observed result, but also on the *probabilities of results that could have been observed* as equation (7.2) makes clear. We can apply this to our medical trial, too. If a standard discrepancy statistic is adopted for computing the p-value, the Binomial design and the negative Binomial design yield different p-values:

$$p_{\text{obs}} = B_{12,0.5}(\{k \geq 9\}) = \frac{299}{4096} \approx 0.073$$

$$\overline{p_{\text{obs}}} = \overline{B_{3,0.5}}(\{n \geq 12\}) = \frac{67}{2048} \approx 0.033.$$

In the remainder of this section, we will examine whether p-values are reasonable measures of evidence or whether they should be replaced by other quantities. We will start with noting some crucial properties of p-values.

First, there is the asymmetry noted in the above Armitage/Berry quote: whereas the rationale for interpreting low p-values is quite clear (namely as evidence against the null hypothesis), it is not clear what a high p-value means – but in any case, it does not necessarily mean evidence for the null. There is an additional issue if p-values are used to compute degrees of severity: a hypothesis can pass a test with quite high severity, but if the result is only slightly modified, it can happen that the same hypothesis is rejected and the alternative passes the test with quite low severity. We saw such an example in the previous chapter (equations (6.4) and (6.5)): 15 successes in a Binomial experiment with $N = 20$ meant rejection of of the null with high severity, 14 successes meant acceptance of the null with low severity. We would prefer a measure of evidence with more continuous transitions, but due to the dichotomous accept/reject character of Neyman-Pearson tests, degrees of severity (and p-values) cannot deliver that.

Second, the choice of the distance function is often a highly non-trivial task. The tenability of a null (or default) hypothesis H_0 is generally evaluated in two types of situations, namely situations with specific alternatives to H_0 and situations where no other hypotheses compete with H_0. In statistical terminology, they correspond to the one-sided and two-sided hypothesis testing problem. In a one-sided testing problem, the distance function d_{H_0} has a specific departure direction measures (towards the alternative hypotheses) whereas in the two-sided problem, no such direction exists. For the one-sided problem, the likelihood ratio (or a monotone function thereof) of H_1 and H_0 $L(H_1, H_0)(x) = P(x|H_1)/P(x|H_0)$ is certainly a good indicator for indicating discrepancy to the null: the greater $L(H_1, H_0)(x)$, the more diverges x from the null, when compared to H_1. Nevertheless, presupposing the existence of a distance function without a specific direction of departure is far from trivial. It is common to invoke the intuition that the less likely a result under the null hypothesis, the more it diverges from the null. Thus, the probability density constitutes a natural choice for a distance function, see again figure 7.1. The p-value is the sum of probabilities of those elements of the sample space that are equally or less likely than the observed value x_0.[14] But we will soon see that this view leads into trouble.

Third, p-values tend to grossly overstate the evidence against the null hypothesis especially when a point null hypothesis is tested against an unspecified alternative. Berger and Sellke (1987) examine the case of normal distributions, testing $H_0 : N(0, \sigma^2)$ against $H_1 : N(\mu, \sigma^2)$, $\mu \neq 0$ by means of several i.i.d. trials, with known σ. If the prior probabilities are impartial ($P(H_0) = P(H_1) = 0.5$) and the observed results significantly diverge from H_0, the p-value is clearly lower than the posterior probability of H_0. More precisely, if the component hypotheses of the alternative H_1 are weighed according to a $N(0, \sigma^2)$ distribution (which is a standard choice), then a p-value of 0.05 implies that the posterior probability of H_0 is *at least* 0.30.[15] In other words, a p-value that is associated with 'strong evidence against the null' does not imply that H_0 is no more sustainable (to say the least), quite to the contrary. This problem occurs because the p-value is not based on the full available knowledge (namely that $x = x_{\text{obs}}$) but only on the knowledge that the observed result is in the set $\{x|d_{H_0}(x) \geq d_{H_0}(x_{\text{obs}})\}$ – the set of all results that have greater or equal discrepancy to the null hypothesis. Therefore it is not surprising that the p-value substantially overestimates the evidence against the null hypothesis.

Even more embarrassing, for a fixed p-value (e.g. $p_{\text{obs}} = 0.05$) and increasing sample size n, the posterior probability of H_0 will eventually converge to 1. Table 7.1, taken from Berger and Sellke 1987, illustrates that phenomenon

[14] Neither is it obvious that there is only one sensible choice for the probability measure in the evaluation of $\{T(X) \geq T(x_0)\}$. If H_0 is a composite hypotheses, Bayesians have to choose between the prior or the posterior distribution of H_0. Some even suggest further calibration. See Bayarri and Berger 2000, Robins et al. 2000.

[15] Theorem 2 in Berger and Sellke 1987, 116.

p-value	$n = 5$	$n = 20$	$n = 100$	$n = 1000$
0.05	0.33	0.42	0.60	0.82
0.01	0.13	0.16	0.27	0.53

Table 7.1: The posterior probability of $H_0 : \mu = 0$ as a function of the sample size n for fixed p-values when $H_1 : \mu \neq 0$ is distributed according to a $N(0, \sigma^2)$-distribution.

for some values of p and n.

The posterior probability converges against 1 because under the given assumptions,

$$P(H_0|X_1, X_2, \dots X_n) = \left[1 + \frac{1}{\sqrt{1+n}} \exp \frac{n\left(\Phi^{-1}(1 - p_{\text{obs}}/2)\right)^2}{2(n+1)}\right]^{-1} \tag{7.3}$$
$$\overset{n\to\infty}{\longrightarrow} 1$$

where Φ is the cumulative distribution function of the standard normal distribution $N(0, 1)$.[16] Thus, a p-value cannot be interpreted regardless of the sample size – the strength of the evidence which they express is quite sensitive to the sample size. In particular, a p-value can be smaller than 0.05 and indicate strong evidence against the null although the observed results favor the null hypothesis over the alternative. This is also called the *Jeffreys-Lindley paradox*.[17] Even if a hypothesis if strongly favored over the alternative by a series of observations (e.g. if $P(H_0|X_1, \dots X_n) \to 1$), the p-value can be extremely low and unfavorable to H_0. This is clearly inadequate. It might be illuminating to track the problem to its sources: the higher the sample size, the more skewed the probability distribution of the sample mean under H_0 and the lower the p-value also for results that only minutely diverge from H_0.

Fourth, p-values depend on the set of *results that could have been observed.* This has some awkward implications for scientific practice. Assume that a malicious experimenter conducts an experiment with stopping rule τ_1. After observing data D, she discovers that the evidence against the null hypothesis is not as strong as she would like to have it. What does she do? She might collect more data, but instead of this cumbersome and potentially expensive activity she has a comfortable shortcut: in her research report, she does

[16] Equation (7.3) easily follows from equation (1.1) in Berger and Sellke 1987, 113.

[17] See Lindley 1957, Jeffreys 1961, Good 1983.

not report the true stopping rule τ_1, but a modified stopping rule τ_2 under which D yields a lower p-value. As readers of a scientific journal, we want to be protected against such tricks. The crucial point is that the malicious experimenter did not manipulate the *data*: she was just insincere about her *intentions* when to terminate the experiment.[18] Using fake data involves considerable risk for an experimenter: if replications fail to reproduce the results, she will lose all her reputation. By contrast, she can never be charged for insincerely reporting her intentions – they are out of reach when double-checking the results. By a post-mortem manipulation of the experimental design, every experimenter has some wiggle room for biasing the results in favor of her preferred conclusion without taking any personal risk. Measures of evidence which have this consequence should not be admissible. The results of experiments and the strength of the observed evidence should be *replicable*, but this condition is odd if the strength of the evidence depends on the experimenter's personal intentions.

Even if all experimenters were completely sincere (certainly an idealizing assumption), caring for stopping rules would severely restrict scientific practice. First, it would be impossible to interpret data that were collected without a definite plan how to conduct the experiment. We would have to conjecture under which circumstances the experiment would have been terminated, but that is very speculative work. Second, stopping rules can be highly contrived and hard to specify: For instance, research funds might be withdrawn or technical problems in conducting an experiment might unexpectedly occur. Then, the experiment would have to be terminated although the proper statistical design did not account for this possibility. In fact, no journal article that reports p-values (and is implicitly committed to the relevance of stopping rules) ever bothers about fine-tuning the stopping rule to the external circumstances under which the experiment was conducted. In principle, one would have to consider all those factors in fixing the stopping rule, but this is practically impossible. Empirical scientists do not take the relevance of stopping rules as seriously as their widespread adherence to the Neyman-Pearson framework of statistical inference suggests. In fact, they have no other choice when they want to maintain ordinary experimental practice.

Fifth and last, it is questionable whether we can at all meaningfully speak

[18] This argument extends a point of Edwards et al. (1963, 239) who stress the irrelevance of the experimenter's intentions to scientific inference.

of 'evidence against H_0' simpliciter, without recourse to explicit alternative hypotheses. This is nothing error or Neyman-Pearson statisticians would do, but p-values are frequently used for this task. The standard example for this kind of reasoning is due to R. A. Fisher (1959) who investigated the hypothesis that the stars are uniformly distributed on the sky, i.e. that the chance that a star is in a particular area of the sky is proportional to the size of the area. Fisher notes that near a particular star (Maia), there are five other stars and believes such an event to be unlikely enough to rule out the hypothesis of uniform distribution. The core of the argument consists in 'Fisher's disjunction':

> "Either an exceptionally rare chance has occurred, or the theory [the null hypothesis, J.S.] is not true."[19]

In other words, results that are very unlikely under the null hypothesis count as strong evidence against the null hypothesis and justify dismissal. Note, by the way, the connection to p-values: The more the actual result diverges from the null hypothesis, the higher the p-value and the stronger, according to Fisher's disjunction, the evidence against the null. But Hacking (1965, 81-82) has convincingly argued that Fisher's disjunction is fallacious. Under the hypothesis of uniform distribution, every possible constellation of stars is equally likely or unlikely. Thus, there are no 'likely chances', but each possible event constitutes an 'exceptionally rare chance'. If Fisher's disjunction were correct, we would always have to reject the hypothesis of uniform distribution as long as there are enough events involved.

An attempt to rescue Fisher's contention from Hacking's objections consists in the interpretation that observing an event that is exceptionally rare *compared to other possible events* is sufficient to rule out the hypothesis at stake. But as Royall (1997, 65-68) has pointed out, such an arguments cannot work either. First, measures of relative unexpectedness involve the likelihood of results that were not observed, thereby depending on the sample space and violating the objectivity conditions. We have already seen that such a dependence poses severe problems for scientific practice – think of the malicious experimenter. But there is a more specific problem, too, as a simple variation of Fisher's example illuminates. We would like to test the hypothesis H_0 that a particular coin is fair. To this end, we take a series of i.i.d. Bernoulli trials and note the observed sequences of heads and tails.

[19] Fisher 1959, 39.

Now, all sequences of heads and tails are equally likely under H_0. Since no sequences is favored over other sequences, it is not possible to reject H_0 on the basis of the relative unexpectedness of the observed result, even if the actual result is 'HHHHHHHHHH' (or a similarly extreme sequence). But this is absurd since by all means, a sequence that consists only of heads seems to provide evidence against the fairness hypothesis (or at least stronger evidence than 'HTTTHTHHTH'). The full data – i.e. the most fine-grained partition of the sample space that we can make – cannot always be the right statistic when testing statistical hypotheses.[20] Hence, we have to individuate the possible observations in a way that avoids the above counterexamples and makes sense of Fisher's disjunction. There, it appears natural to count only the number of heads or tails that occurred in the trial, because we believe that the order of heads and tails in the sequence does not matter at all and because this is the most compressed form in which we can represent the informational content of the data. Only if the data are compressed to a minimal sufficient statistic, 'equivalent' sequences as 'HHHHHTTTTT' and 'TTTTTHHHHH' correspond to the same observation and counterexamples of the above type are avoided. Now, since the number of heads is a minimally sufficient statistic with regard to the propensity ϑ of the coin to fall heads, we can say that 'HHHHHHHHHH' is an exceptionally rare chance with regard to the fairness hypothesis $\vartheta = 0.5$. Thus, our best candidate for explaining Fisher's notion of an 'exceptionally rare chance' seems to be an event that is *relatively improbable compared to other events*, where events correspond to possible values of a statistic that is minimally sufficient with regard to the parameter of interest ϑ. In other words, there is no exceptionally rare chance as such – any such chance is relative to the choice of a statistic that determines *the way in which it is exceptional.*[21]

But actually, introducing a minimally sufficient statistic into the explication of 'exceptionally rare chance' introduces implicit alternative hypotheses, too. When relativizing unexpectedness to a parameter of interest, we have committed ourselves to a specific class of potential alternative hypotheses – namely those hypotheses that correspond to the other parameter values. In

[20] See Seidenfeld 1979a, 80. Seidenfeld also discusses Fisher's disjunction, but under the (equivalent) label of 'significance tests'.

[21] In the present case, it appears at superficial sight that there can be only one parameter of interest. But some model families have two or more parameters, e.g. mean and variance in the case of the normal distribution. For instance, the sample mean is minimally sufficient for the population mean, but not for the population variance.

our example, this was $\vartheta \neq 0.5$. When applying Fisher's disjunction, we do not judge the tenability of H_0 'in general', without recourse to a specific parameter or comparison to alternatives – we always examine a certain way the data could be surprising. Thus, when applying Fisher's disjunction, we are asking specific questions about a parameter as 'why that value of ϑ rather than another one?'. The choice of the minimally sufficient statistic required to apply Fisher's disjunction reveals a class of intended alternatives. This has some general morals: what makes an observation evidence against a hypothesis is not its low probability under this hypothesis, but its low probability compared to an alternative hypothesis. An improbable event is not evidence against a hypothesis per se, but

> "[...] what it does show is that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability [...] you will be very much more inclined to consider that the original hypothesis is not true."[22]

To summarize: Fisher's disjunction and the inference from relatively unlikely results to evidence is caught in a dilemma: Either we summarize various possible results into one equivalence class. Then the choice of the test statistic reveals implicit alternatives to which the hypothesis is compared. Or we apply Fisher's disjunction based on the most fine-grained available partition of the sample space (i.e. the space of possible observations). But then, some hypotheses (as the fairness hypothesis in the coin flip example) cannot be tested at all with the help of Fisher's disjunction. Thus, evidence has to be a comparative concept in the sense that evidence *against* a hypothesis is always evidence *for* another hypothesis.

So far we have presented a negative characterization of evidence and arguments why error probabilities and degrees of severity cannot figure as an adequate measure of evidence. For establishing how a measure of evidence should look like, it is now time to introduce some more formal constraints on measures of evidence. So far, it has become clear that a measure of evidence must be comparative – it depends on the observed data $x \in X$ and two competing hypothesis $H_0, H_1 \subset \Theta$. For reasons of simplicity, we restrict ourselves to point hypotheses, i.e. H_0 and H_1 correspond to $\vartheta_0 \in \Theta$ and $\vartheta_0 \in \Theta$ so that $Ev : X \times \Theta \times \Theta \to \mathbb{R}$. In his penetrating discussion of the

[22] William S. Gosset ("Student") in private communication to Egon Pearson, quoted in Royall 1997, 68.

subject matter, Subhash Lele (2004) suggests a list of criteria from which I pick the central ones:

Antisymmetry The evidence which x gives for ϑ_0 and against ϑ_1 is the negative of the evidence x gives for ϑ_1 and against ϑ_0: $Ev(x, \vartheta_0, \vartheta_1) = -Ev(x, \vartheta_1, \vartheta_0)$. In particular, $\forall \vartheta : Ev(x, \vartheta, \vartheta) = 0$.[23]

Reparametrization Invariance Reparametrizing a parameter space, e.g. with the help of a map $\psi : \Theta \rightarrow \Psi$ should not modify the strength of evidence because such a reparametrization amounts to a mere mathematical manipulation.

Data Transformation Invariance If $g : X \rightarrow Y$ is a bijective transformation of the data, x and $g(x)$ induce the same evidence function (relative to the distributions which ϑ_0 and ϑ_1 induce on Y): $Ev(x, \vartheta_0, \vartheta_1) = Ev(g(x), \overline{g}(\vartheta_0), \overline{g}(\vartheta_1))$.

Maximization On average, the value of Ev is maximized at the true parameter value: $\mathbb{E}_{\vartheta_0}[Ev(x, \vartheta_0, \vartheta_1)] < 0 \, \forall \vartheta_1 \neq \vartheta_0$.

Laws of the Large Numbers In the long run, Ev converges against its expectation. In other words, P_{ϑ_1}-stochastically,

$$1/n[Ev(x, \vartheta_0, \vartheta_1) - \mathbb{E}_{\vartheta_0}[Ev(x, \vartheta_0, \vartheta_1)]] \rightarrow 0.$$

By means of these criteria and a number of other reasonable constraints, Lele is able to show that for two point hypotheses, the log-likelihood ratio

$$\log \frac{P(x|\vartheta_1)}{P(x|\vartheta_0)}$$

emerges as the optimal evidence function, up to normalization. Such an understanding of evidence opens the way to representing data as evidence and as a basis for scientific inference. A log-likelihood ratio greater than 2 (-2) then counts as strong evidence and a log-likelihood ratio greater than 4 (-4) as very strong evidence whereas the are $[-1, 1]$-area merely indicates weak evidence. There are a lot of further fruitful applications – for instance,

[23] In the special case that ϑ_0 and ϑ_1 exhaust the hypothesis space, antisymmetry reduces to the symmetry constraint $c(H, E, K) = -c(\neg H, E, K)$ in Crupi et al. (2007).

Royall (2000) gives upper bounds for the probability of observing misleading evidence in a statistical experiment which can be phrased as bounding $P_{\vartheta_1}(Ev(x, \vartheta_0, vt_1) > K)$, K very large. This procedure is analogous to the pre-experimental specification of error probabilities in a error-statistical framework.

It is easy to see that the log-likelihood ratio merely depends on the likelihood function of the data and thus satisfies the Likelihood Principle, leaving no room for subjective distortion. Therefore, they are sufficiently objective in order to be suitable for scientific communication. These observations do not change when H_0 or H_1 are composite hypotheses. Admittedly, prior probabilities appear in computing

$$P(x|H_1) = \int_{\vartheta_1 \in H_1} d\vartheta_1 P(\vartheta_1|H_1)\, P(x|\vartheta_1).$$

So testing becomes more subjective for stating the evidence for and against composite hypothesis. But still, dissent about the strength of evidence can always be tracked to different assignments of prior probabilities. Furthermore, objective and noncommittal priors (Jeffreys 1961) can often be used to solve that problem. Typically, likelihood ratios are generalized into Bayes factors:

$$B(H_1, H_0, x) := \frac{P(H_1|x)}{P(H_1)} \frac{P(H_0|x)}{P(H_0)} = \frac{\int_{H_1} P(\vartheta)\, P(x|\vartheta)\, d\vartheta}{\int_{H_0} P(\vartheta)\, P(x|\vartheta)\, d\vartheta}. \tag{7.4}$$

Despite the inherent subjective component in (7.4), Bayes factors provide useful measures of evidence. Think again of the example at the beginning of this chapter: In a repeated Bernoulli trial, we test $H_0 : \vartheta = 0.5$ against $H_1 : \vartheta > 0.5$. Assume that H_1 itself is uniformly distributed, i.e. according to H_1, $\vartheta \sim U(0.5, 1)$. After twelve tosses, nine recoveries and three failures have been observed. This means that the logarithmic Bayes factor in favor of H_1 over H_0 is

$$\begin{aligned} \log B(H_1, H_0, x) \ &:= \ \log \frac{P(H_1|x)}{P(H_1)} \frac{P(H_0|x)}{P(H_0)} = \frac{\int P(\vartheta|H_1)\, P(x|\vartheta, H_1)\, d\vartheta}{\int P(\vartheta|H_0)\, P(x|\vartheta, H_0)\, d\vartheta} \\ &= \ \log \frac{2 \int_{0.5}^{1} d\vartheta \begin{pmatrix} 12 \\ 9 \end{pmatrix} \vartheta^9 (1-\vartheta)^3}{\begin{pmatrix} 12 \\ 9 \end{pmatrix} 0.5^{12}} \\ &= \ 0.437. \end{aligned}$$

The Bayes factor indicates evidence for H_1 against H_0, but it is far from indicating very strong or overwhelming evidence. Thus, the Bayes factor (a generalization of the likelihood ratio) allows for a more fine-grained and meaningful representation of statistical evidence than the result of Neyman-Pearson tests ('accept/reject') or the p-values/degrees of severity whose problems we have realized by now. Despite their subjective components, likelihood rations and Bayes factors are superior to p-values. Before bringing to bear the results on the debate around design and stopping rules, I would like to make some remarks on the epistemological function of p-values.

7.3 P-values revisited

Despite the frequently found misinterpretations, the ubiquitous use of p-values in the empirical sciences suggests that there is something p-values are good for. This brief section presents another application of p-values which is more adequate than their (mis)use as measures of evidence. Since more than two decades, statisticians have been researching on the connection between p-values and Bayesian measures of evidence. Indeed, there is a compatibility result when p-values are used in testing a hypothesis against specific alternatives so that a specific direction of departure from the null hypothesis is distinguished. This situation is quite different from the one we encountered in the previous section where point null hypotheses were tested against unspecified alternatives. There, p-values performed quite poorly and overstated the evidence against the null, inter alia. In the new type of situations, the situation is not as bad as there. Again, we focus our discussion on a normal distribution $N(\mu, \sigma^2)$ with known variance σ^2 and unknown mean μ. The rivalling hypotheses are $H_0 : \mu \leq 0$ and $H_1 : \mu > 0$.

Casella and Berger (1987) show that the p-value of with regard to H_0, $\overline{X}_n$ and $T(x) := x$ provides a *lower bound for the posterior probability of* H_0, taken over a certain class of prior densities π that assign equal weight to both hypotheses. In mathematical terms,

$$\inf_{\pi} P(\mu \leq 0|x_0) = p_{\text{obs}}(x_0) := P(X \geq x_0|\mu = 0), \tag{7.5}$$

or equivalently,

$$1 - p_{\text{obs}}(x_0) = \sup_{\pi} P(\mu > 0|x_0) \tag{7.6}$$

(Theorem 3.2. in Casella and Berger 1987, 108).[24]) In other words, under suitable (and 'impartial') prior assignments, the null hypothesis is at least as likely as the p-value indicates, and in some cases, this bound is almost attained. Hence, p-values come much closer to the posterior probability of H_0 than in the case of testing the point null hypothesis $\mu = 0$ against the unspecified alternative $\mu \neq 0$. This result also explains why Bayesian posterior probabilities and p-values are often conflated in statistical practice.

In the one-sided testing problem, the p-value $p_{\text{obs}}(x_0) = P(X \geq x_0 | \mu = 0)$ sums up the probability of those values where the evidence in favor of H_1 and against H_0, as measured by the Bayes factor, is greater than at the actually observed value x_0. In a similar vein, under a suitably narrow class of prior distributions π and alternatives H_1, p-values can be calibrated as to provide lower bounds on Bayes factors (see Sellke et al. 2001):

$$\inf_{\pi} B(H_1, H_0, x) = -e p_{\text{obs}}(x) \log p_{\text{obs}}(x). \tag{7.7}$$

Hence, we see how p-values can be calibrated as to provide lower bounds for the strength of the evidence and the posterior probability of the null. In practice, this can be very useful: instead of a cumbersome and computationally expensive Bayesian analysis, a quickly performed computation of the p-values gives a rough idea of whether the null hypothesis is severely shaken by the data. The p-value is easy to calculate and avoids careful deliberation about prior probabilities etc. Therefore, computing p-values can be a quick and dirty way to make further calculations superfluous. For instance, if the p-value is greater than 0.1, we know that the null hypothesis has *at least* a probability of 0.1 so that it remains a serious candidate. Rightfully, it is often stressed that the use of p-values has merely auxiliary character; as soon as a full analysis is possible, they cannot play any role. So, although p-values can give a rough idea about a the evidential content of the data, they are only preliminaries to the computation of the actual strength of evidence or a final judgment on the tenability of a hypothesis. In the latter case, this is particularly salient because they merely provide lower bounds for the posterior probability of the null hypothesis (respectively the evidence against the null) where an upper bound would be required. Hence, although no scientific report should cite the observed p-value in favor of rejecting the

[24] Casella and Berger even derive this result for any distribution family that is indexed by μ that is (1) symmetric around zero and (2) has a monotonously increasing likelihood ratio.

null hypothesis (as it is often done, unfortunately), working with p-values remains practically useful and has a heuristic value.[25]

7.4 Putative reductios

The preceding arguments have driven us towards the evidential, post-observational irrelevance of experimental design: The measures of evidence to which our deliberations led us satisfied the Likelihood Principle and thus the Stopping Rule Principle, too. Nonetheless, error-statistical intuitions about stopping rules are resilient – somehow they seem to matter in spite of all arguments to the contrary.

Defenders of the inferential role of experimental design usually try to beat the Bayesians in their own game. Bayesians are allegedly unable to detect that an experiment has been designed as to *reason to a foregone conclusion.* Imagine a football match between Neyman-Pearson Wanderers and Bayesians United. The rules of the game are slightly amended: Neyman-Pearson Wanderers are granted the right to terminate the match at any moment they wish. We are told the result of the match: Neyman-Pearson Wanderers won 1:0. Now the coach of Bayesians United, due to his Bayesian conviction, seems to be unable to note that the design of the match was unfair to his team and favored the opponents. Since a Bayesian only listens to the data, the unfair experimental design falls out of his inferential scheme and the coach of Bayesians United has no incentive to complain, or so error statisticians argue.

We believe, however, that this objection (and similar ones) can be rebutted. First, it will be shown that the error-statistical reconstruction of the Bayesian position is not entirely fair. Hence, the error-statistical counterexamples to the Bayesian position are not compelling. Second, Kadane et al. (1996a, 1996b) have shown that it is not possible to reason to a foregone conclusion and to discredit a true hypothesis 'for free'. Posterior probabilities of a hypothesis can not be arbitrarily manipulated with the help of a stopping rule. If we stop an experiment if and only if the probability of a hypothesis falls below a certain threshold, there will be a substantial chance that the

[25] Recall however, that these results hold for p-values with a specified direction of departure. When no such direction is specified, p-values grossly overstate the evidence against the null, and their interpretation becomes much more difficult, as it was shown by Berger and Sellke 1987.

experiment will never terminate.

Let's start with the putative error-statistical reductio of Bayesianism. Bayesians typically argue that Bayesian measures of evidence (Bayes factors, log-likelihood ratios) are superior to error-statistical measures so that they do not need to care for error-statistical arguments that point to the relevance of stopping rules. Error statisticians then try to show that Bayesian measures of evidence, e.g. the posterior probability, are also affected by the stopping rule. Consider the following example (due to Mayo and Kruse 2001): We sample from a normal distribution with variance 1, $N(\mu, 1)$. According to the null hypothesis, the mean of the distribution is 0, i.e. $H_0 : \mu = 0$ which is tested against the unspecified alternative $H_1 : \mu \neq 0$. Now, we decide to take samples from the sequence $(X_n)_{n \in \mathbb{N}}$ (X_k are i.i.d. standard normals) until the inequality

$$\left|\overline{X}_n\right| := \left|\frac{1}{n}\sum_{k=1}^{n} X_k\right| \geq \frac{2}{\sqrt{n}} \tag{7.8}$$

is satisfied, i.e. until the absolute value of the sample mean is greater than $\frac{2}{\sqrt{n}}$. For any n, $P(\left|\overline{X}_n\right| \geq 2/\sqrt{n}|H_0) \leq 0.05$. Denote this stopping rule by the letter τ. It is straightforward to show that τ is almost certainly bound to terminate, regardless of whether H_0 is true or not. Therefore, τ terminates at a certain N. For any such N,

$$\begin{aligned} P(H_0|\overline{X}_N) = \frac{P(H_0)P(\overline{X}_N|H_0)}{P(\overline{X}_N)} \quad &< \quad P(H_0) \\ \Leftrightarrow P(\overline{X}_N|\neg H_0) > P(\overline{X}_N|H_0) \end{aligned}$$

as it can be shown by a straightforward calculation. In other words, the prior probability of H_0 exceeds the posterior probability if and only if the observed outcome $\overline{X}_N$ is more likely under the alternative hypotheses than under the null.[26]

For example, if the prior distribution of the mean μ under $H_1 := \neg H_0 = \mu \neq 0$ is uniformly distributed, i.e. $\mu \sim U[-\infty, \infty]$, the prior probability will exceed the posterior probability *regardless of the precise time point when the experiment terminates.* Hence, for such priors, the stopping rule ensures that our degree in belief in H_0 is going to decline. In other words: Given the stopping rule, it is almost certain that we will end with a posterior

[26] $\overline{X}_N$ is a sufficient statistic. Then, the sufficiency principle asserts that we do not need to know further details about the data, see Birnbaum 1962.

probability of H_0 lower than the prior probability, and this should trouble a Bayesian who asserts the irrelevance of stopping rules. Stopping rules allow to manipulate the posterior probability and thus bias the rational degrees of beliefs. The outlined experiment artificially decreases the rational degree of belief, regardless of the precise outcome.

However, the example only works if the alternative hypotheses are uniformly distributed over the real line. Such an 'improper' prior distribution is mathematically tractable, but strictly spoken, it is not admissible because the probabilities do not add up to 1. This is the reason for the decrease in the posterior probability. In fact, we have already shown that the test works the other way round for proper priors: Assume that the parameter values $H_1 : \mu \neq 0$ are distributed according to $N(0, 1)$ and assume furthermore that $P(H_0) = P(H_1) = 0.5$. Then, the stopping rule

$$\begin{aligned} \tau : X^{\mathbb{N}} &\rightarrow \mathbb{N} \\ (X_N)_{N\in\mathbb{N}} &\mapsto \min_{N\in\mathbb{N}}(\overline{X_N} \geq 2/\sqrt{N}) \end{aligned}$$

amounts to terminating at the fixed p-value 0.05. For this case, the Jeffrey-Lindley paradox applies and for increasing n, *the posterior probability of H_0 actually goes to 1.* Thus, for increasing n, the p-values remains low, but the posterior probability increases, as witnessed by table 7.1. Hence, the argument against the Bayesian neglect of stopping rules dissolves for the case of proper priors.

But why did the argument work for improper priors? When improper priors are used with an increasing number of samples, most of the posterior mass will concentrate on the alternatives close to the observed sample mean: for each observed sample mean, there is a corresponding 'optimal' hypothesis which will be favored by the data. The comparison to the finite case illuminates the problems of such priors. Assume that the alternative hypothesis consists of a finite set of mean values $\mu \neq 0$. Then the above stopping rule does not necessarily lead to a lower posterior probability because the cutoff point $\frac{1}{2}/\sqrt{N}$ will, as N increases, be closer to $\mu = 0$ than to the available alternatives, thus favoring H_0 over H_1. The effect on which Mayo and Kruse's example builds is random sampling variation – even if $H_0 : \mu = 0$ is the true distribution, almost certainly a wrong distribution that is sufficiently close to H_0 will be favored over H_0. But if the space of hypotheses is finite or if proper priors are used – and that are the realistic cases – the argument is no more applicable. Hence, the attempt to beat the Bayesian according to his own standards, fails.

More generally, define a stopping rule τ_2 so that it terminates as soon as the posterior probability of H_0 falls under a certain value q: Then, Kadane, Schervish and Seidenfeld show that

$$\begin{aligned} P(\tau_2 < \infty) &\leq \frac{P(H_0)}{1-q} < 1 \\ P(\tau_2 < \infty | H_0) &\leq \frac{(1-P(H_0))}{P(H_0)} \frac{q}{1-q} < 1 \end{aligned}$$

which implies that we cannot manipulate posterior probabilities without exposing ourselves to the risk of infinite sampling ($P(\tau_2 = \infty) > 0$). In other words, the above equations illustrate the impossibility of 'reasoning to a foregone conclusion'.[27] If a certain posterior probability has to be achieved, there is a non-trivial chance that sampling will continue forever. Thus, posterior probabilities cannot be deliberately manipulated.

7.5 A decision-theoretic perspective

Finally, we have to integrate our conclusions into a decision-theoretic framework and to defend it against attempts to render it incoherent. Furthermore, we have to explain in how far experimental design is scientifically relevant and why it often appears to be evidentially relevant, too.

A statistical decision rule is a function from the set of possible observations (which depends on the particular experimental design) to a set of actions, e.g. acceptance or rejection of a hypothesis. Abraham Wald's (1950) criterion of *admissibility* demands that no statistical decision rule be *(weakly) dominated* by another one. A decision rule is dominated if it selects an inferior outcome in all circumstances, regardless of 'nature's choice' with respect to the truth or falsity of the available hypotheses. Choosing such a rule would decrease the payoff uniformly and thus count as irrational behavior, like in game theory. This is not only important for Neyman-Pearson statisticians, but also for Bayesians, because only admissible decision rules minimize the expected risk relative to some prior distribution. Inadmissible rules are always inferior. Now we can see what has gone wrong in the football match: By

[27] See Kadane et al. 1996b, S283.

agreeing to the modified rules, Bayesians United have committed themselves to an inadmissible decision rule which is weakly dominated by the normal rules. Bayesians United have put themselves in a worse position, regardless of the actual course of events and have thus acted irrationally. The need for admissible decision rules illustrates why Bayesians should pay attention to experimental design, too. When we claim the 'intuitive' relevance of stopping rules and design, we normally have such considerations in mind.

But what about the post-observational evaluation of the football match? Didn't the unfair setup bias the available evidence? A positive answer is tempting, but wrong. A famous result of statistics, the Rao-Blackwell-theorem, helps us to see why. Strangely enough, it is not cited in the debate about stopping rules. The theorem guarantees that no information beyond the data (more precisely, a sufficient transformation of the data) is able to improve a post-experimental decision rule and to lower the risk associated with a decision for or against a hypothesis.[28] In particular, any statistical decision rule that sensitively depends on the stopping rule can be improved by eliminating the stopping rule dependency and conditioning on the data only. Thus, such rules are weakly dominated and inadmissible. Information about the experimental design must be irrelevant to a rational decision that is made after seeing the results. Thus, the Bayesian position on stopping rules can be saved from attempts to render it incoherent. In the football example, the *evidence* or insight which the match delivers about the abilities of the teams is not affected by the reasons for terminating the match. Note that the match could have been terminated due to a sudden rainfall, too. In a decision-theoretic framework, any post-mortem decision on which team is the better one is just a function of the match observations, as a consequence of the Rao-Blackwell theorem. Thus, the evidential irrelevance of the experimental design agrees with the principles of rational choice.

Furthermore, the choice of an experimental design may help to minimize costs when single trials are expensive, e.g. in medicine. Assume that the surveillance of every trial person costs a substantial and fixed amount of money. Then, we want to avoid an experimental design which advises to continue sampling when the evidence in favor of a hypothesis is already overwhelmingly strong. This might happen in a trial with fixed sample size, for instance. Making additional trials would imply unnecessary costs so that

[28] This holds for a convex loss function – an assumption that is usually satisfied.

Utility/Design	zero costs		Binomial design		neg. Bin. design	
true hypothesis	H_0	H_1	H_0	H_1	H_0	H_1
decision for H_0	100	0	88	-12	94	-6
decision for H_1	0	50	-12	38	-12	38

Table 7.2: A sequential Bernoulli trial of $H_0 : p = 1/2$ against $H_1 : p = 1/4$. Each sample is either a success or a failure, p giving the success probability. The table shows the utility of a decision for H_0/H_1 (rows) when H_0/H_1 is the case (columns). The left table neglects the experimental costs, the other tables discount the gains by the expected sample size, each sample having a fixed cost of 1. In the middle table, the sample size is fixed to $N = 12$ (Binomial design), whereas the right table fixes the total number of successes to $K = 3$ (negative Binomial design).

there is, at last, a connection between experimental design and rational behavior. For this reason, the design of sequential trials has developed into an art of its own (see Wald 1947, Armitage 1975). The power of the chosen decision rule to discern the true hypothesis has to be weighed against the costs of the specific design, usually measured by the expected number of samples. We have to optimize the number of sequential trials relative to the gain that a correct decision promises. Table 7.2 illustrates this point and compares three utility matrices: in the left one, costs are zero whereas in the latter two matrices, the gains are discounted with the expected number of samples under two different designs (see appendix B for details). Notice that the negative Binomial design is more efficient than the Binomial design which is not admissible. Thus, it is a misunderstanding that Bayesians do not care at all for experimental design: the loss function is crucially affected by the possible outcomes, viz. the experimental design. But the relevance of experimental design is purely a pre-observational one and does not affect statistical decisions once the data have been observed.[29]

This point can be generalized: A Neyman-Pearson hypothesis test corresponds to a statistical decision rule. Assume that we have two Neyman-Pearson tests with different characteristic error probabilities $< \alpha, \beta >$ and $< \alpha', \beta' >$. The admissibility criterion implies that no test should be weakly dominated by the other one (i.e. $\alpha \leq \alpha'$ and $\beta \leq \beta'$). If both tests are admissible, it will depend on the prior probabilities and the utility matrix

[29] Experimental design is also important for *secure reasoning* and robust inference (Staley 2004). Exploring this connection would, however, go beyond the scope of this article.

which of the two tests we should prefer. Here we spot the connection between admissibility as a rationality constraint for statistical decision rules and the significance of error probabilities: once the priors and the utility matrix are fixed, the goodness of a decision rule can be assessed by means of the error probabilities (details omitted). It is therefore a misunderstanding, though a widespread one, that Bayesians are indifferent to error probabilities – and I believe that much of the heat in the Bayes vs. Neyman-Pearson debate is owed to that wrong perception of Bayesianism. Error probabilities play a crucial role in the pre-experimental assessment of a decision rule and experimental design is important for cost optimization in sequential trials. Philosophers of statistics in the Neyman-Pearson tradition (e.g. Mayo 1996, Mayo and Spanos 2006) are right to point that out, but when they go beyond this concession and base a theory of statistical inference on error probabilities, p-values or degrees of severity, they are wrong, as argued above. Admittedly, many scientists might be reluctant to specify prior degrees of belief that are required for a Bayesian analysis because 'scientific objectivity' might get lost. But prior beliefs are relevant for rational statistical decisions and scientific inference, so it is only fair and honest not to neglect them and to make them explicit.

7.6 Summary

The debate about the relevance of experimental design and stopping rules is blurred by the lack of clarity which kind of relevance is meant. Equivocation and confusion results. Moreover, the debate is characterized by a mutual deadlock. To resolve it, we have argued that the post-experimental, evidential relevance of experimental design should be denied, for foundational reasons as well as for the needs of empirically working statisticians.

In order to bring out that argument, this chapter has elaborated the connections between the dispute about measures of evidence and the post-observational relevance of experimental design. The conflicting positions in the design debate also take different stands on measures of statistical evidence. I have pointed out that Neyman-Pearson theory of statistical testing is not able to deliver an adequate post-experimental measure of evidence. The more comprehensive error-statistical framework tries to integrate Neyman-Pearson theory into a philosophy of statistical inference that also provides post-observational measures of concordance and dissent. However, such measures (as degrees of severity and p-values) fall prey to numerous objections. The three most important ones were: (1) p-values and degrees of severity often overstate available evidence (2) some implications (as the dependency on the sample space, i.e. the experimenter's intentions) severely restrict scientific practice and (3) it is not possible to measure the strength of evidence against or for a hypothesis simpliciter as the (mis)use of p-values suggests. On these grounds, we have aimed for a comparative measure of evidence that is immune to the above objections, and a number of reasonable adequacy conditions has led us to accepting the (logarithmic) likelihood ratio and its generalization, the Bayes factor, as a suitable measure of evidence. Consequently, p-values cannot count as genuine measures of evidence, but in certain circumstances, they can still figure as heuristic devices for giving mathematical bounds on measures of evidence and posterior probabilities. The failure of p-values and degrees of severity implies that error-statistical inferences (and Neyman-Pearson tests) are caught in the dilemma of either denying the significance of post-experimental data analysis or drawing conclusions which do not help the practicing scientist, as 'this procedure had a reliability of 0.95'. Reiterating a point from the main part of the chapter, scientists do not want to know how frequently such a procedure is successful, but how much *actual confidence* they should lend to a hypothesis which was accepted by a statistical test.

A corollary of the above consists in the post-observational irrelevance of stopping rules and experimental design. This position can be coherently integrated into statistical decision theory, despite the Neyman-Pearson counterarguments. Nonetheless, we have also emphasized the general scientific relevance of experimental design. This is illustrated by the need for admissible statistical decision rules and cost minimization in sequential trials. Experimental design is indispensable for scientific inference, though in a more narrow sense than Neyman-Pearson statisticians believe.

There is a more general moral, too, which was mentioned in passing: error-statistical inference is not as objective as its proponents believe. The charge of subjectivity raised against the Bayesians turns out to be a boomerang since the evidential relevance of the sample space and the stopping rule introduces a source of subjective bias and distortion. Even worse, Bayesians are *honestly subjective* whereas error statisticians are subjectivists who pretend to be objective. Bayesianism is a subjective theory of inductive inference, but its adherence to the Likelihood Principle also makes clear that it is able to give an objective theory of *evidence*. Moreover, the Bayesian's subjectivity can always be tracked to its sources, namely the assignment of prior probabilities, whereas the subjectivity of the error statistician is not accessible to open discussion (see the example on page 180). By eschewing the use of rational credences and subjective probabilities in statistical inferences, the error-statistical approach deprives itself of the ability to answer the central questions in data analysis.

Chapter 8

Summary and Conclusions

This book has compared and discussed various ways to model inductive inference and to capture confirmation, support and evidential relevance. At the beginning, however, we had to connect confirmation theory – the theory of valid inductive inferences – with the principal problem of justifying inductive inferences (chapter one). Famously, David Hume showed that the standard argument for the validity of induction which invokes the past successes of inductive reasoning is itself an inductive argument so that the justification becomes circular. Indeed, the search for a meta-principle that justifies the induction principle opens the door to the classical justification trilemma: regress (how can we justify such a meta-principle?), circularity, or abort. A possible reply that abandons the search for such meta-principles consists in epistemic reliabilism: We are justified in making inductive inferences as long as they are *factually* reliable and produce more true than false beliefs. Nelson Goodman (1983) pointed out, however, that the crucial question is not *whether* induction is a reliable principle but *which kind* of induction is reliable and makes more correct than incorrect predictions. The famous 'grue'[1] example illuminates this point – from a purely logical point of view, the inference to 'all emeralds are grue' conforms to the principles of induction as much as the inference to 'all emeralds are green'. Both hypothesis are inconsistent with each other, and we strongly feel that only the 'green' hypothesis is inductively supported. Goodman convincingly shows that confirmation-theoretic attempts to capture the relevant difference between the 'green' and the 'grue' predicate in purely formal terms are doomed. The task of discern-

[1] Recall that something is grue if and only if (1) it is green and has been examined in the past (before t_0) or (2) it is blue and has not been examined in the past.

ing inductively projectible predicates falls outside the scope of confirmation theory. So confirmation theory has to presuppose a set of projectible predicates. Only then, it can extract valid rules of induction from our inductive practice and formulates those rules in a way that they serve in turn as a corrective for inductive practice. A proper confirmation theory should mirror paradigmatic cases of confirmation in science as well as provide guidelines for assessing the impact of evidence on theory and vice versa. Such a project is sensible even in the face of the principal unresolved difficulties with inductive inference because very often, we know that the predicates at hand are lawlike and confirmable. On the other hand, we are still uncertain how to describe the relation between theory and evidence. In so far as we pursue a *formal* account of confirmation, our project is *explicative* – the vague and informal concept of confirmation is replaced by a similar, but fruitful and exact concept (see Carnap 1950, §3).

Quite obviously, confirmation in science is a very versatile and multifaceted concept. Kepler's laws of planetary motion are confirmed by observations of planetary motions on the nocturnal sky and the wave nature of light is confirmed by experimental scrutiny in Young's double-slit trial. Darwin found evidence for his theory of evolution by excavating fossils in South America. Eddington successfully checked the predictions of Einstein's General Theory of Relativity during the 1919 eclipse. Statistical regularities as Mendel's laws of inheritance are confirmed in controlled, randomized experiments with a large number of trial plants. All those cases are very diverse so that there cannot be a single theory of confirmation – rather we need several accounts of confirmation, corresponding to the different needs of empirical scientists. Two grand traditions characterize the field of confirmation theory: the qualitative and the quantitative (usually Bayesian) tradition. Proponents of qualitative accounts as Clark Glymour (1980a) deny that quantitative, probabilistic accounts are applicable to a wide range of cases of confirmation in the history of science because the probabilistic superstructure which those accounts impose fails to illuminate the intricate structure of confirmatory arguments. On the other hand, purely qualitative accounts are not able to model statistical regularities and to quantify which hypotheses are better confirmed than others. All accounts of confirmation agree, however, that background assumptions play a crucial and autonomous role in assessing the relationship between theory and evidence. Hence, confirmation should be modeled as a three-place rather than as a two-place relation.

Since the mutual objections to qualitative and quantitative confirmation theory are all eligible to a certain degree, both approaches have to be discussed in more or less detail in order to elicit in how far they are able to give a convincing explication of inductive support. I started with qualitative accounts of confirmation (chapter two) – a field that is characterized by two major approaches: inductivist approaches as Hempel's satisfaction criterion and the hypothetico-deductive approach which goes back to Popper's model of prediction, test and empirical corroboration. Hempel's satisfaction criterion suffers under a lot of technically-minded objections (pp. 27-29), but that was not even the main problem. In his [1945] 1965, Hempel gives a convincing analysis of the raven paradox, pointing out that tacitly introduced differences in the background knowledge are responsible for misleading intuitions and the paradoxical appearance of the problem. Indeed, relations of confirmation and support are in general sensitive to adding surplus background knowledge. This is one of the main differences between deductive and inductive logic – contrary to deductive inference, inductive inference is not monotonous. Unfortunately, Hempel does not integrate that insight into his own criterion of confirmation – the satisfaction criterion respects monotony. Thus, it does not only fail to give a convincing reading of the raven paradox, it also fails whenever relations of confirmation are changed by adding substantial background information.

The second tradition – hypothetico-deductivism or briefly, H-D confirmation, – is based on the idea that theories make predictions with the help of auxiliary hypotheses and that those theories are confirmed just in case those predictions are indeed observed. In other words, this model of confirmation describes how hypotheses successfully survive experimental tests where the predictions are derived from the hypothesis under test plus some hypotheses in use. This account of confirmation has, however, longstanding problems to correctly describe evidential relevance as the *tacking paradoxes* make clear: If E H-D confirms H, then E also H-D confirms $H.X$ for an arbitrary X because E is still logically entailed by $H.X$. But in general, E is not relevant at all to X. Several attempts to solve these problems failed. Only in more recent works by Schurz (1991) and Gemes (1993) satisfactory answers have been found. Nevertheless, the latter suggestions are deficient in an important respect, too, namely the confirmation of conjunctively composed hypotheses. Schurz's and Gemes's models of H-D confirmation often imply that if E_1 confirms H_1 and E_2 confirms H_2 then $E_1.E_2$ confirms $H_1.H_2$, too. In a wide range of cases where scientific confirmation is modeled, such a reasoning is misguided, as demonstrated in several examples. In particular, the intuition that *instances* figure centrally in the confirmation of a hypothesis gets lost in that scheme of inference. Therefore I make a new proposal that connects the confirming power of instances with the hypothetico-deductive tradition: the falsificationist criterion of confirmation.[2] Then, I show that this falsificationist proposal resolves the tacking paradoxes as well as the problems with the confirmation of composite hypotheses, unlike any other proposal. Due to its simplicity and its combination of instantial and deductivist views about confirmation, I believe the falsificationist criterion to be the most advanced and accomplished criterion for qualitative confirmation.

Scientists often want to confirm or to refute theories as a whole – an endeavor that is also motivated by the theory-ladenness of observation and the lack of theory-independent neutral background knowledge (Kuhn 1962). Thus, there is demand for a model of theory confirmation that supervenes on relations of evidential relevance between parts of the theory and observed evidence. Such a model, called *bootstrap confirmation*, has been developed by Clark Glymour (1980a): theories are confirmed by deductive moves from evidence plus parts of the theory to other parts of the theory. However, several technical objections (Christensen 1983, 1990) have cast doubt on the feasibility of Glymour's approach – the relationship of evidential relevance is not properly explicated. Two principled answers are possible: Either we content ourselves with bootstrap confirmation as a model of *coherence* between theory and evidence, or we make some technical modifications, replacing Hempel's satisfaction criterion in the bootstrap account by the falsificationist criterion. I pursue the latter road and show how the falsificationist criterion is able to rescue bootstrap confirmation and to counter Christensen's objections for a suitably modified account of bootstrapping. Thus, we obtain a viable account of bootstrapping which could be expanded in future work.

Nevertheless, all accounts of confirmation have to defend themselves against the Duhem-Quine objection (chapter three). Duhem (1914) elaborated that falsification of a hypothesis is only as reliable as the auxiliary theories used in the falsification of that hypothesis. Quine (1961) extended this observation to the epistemological tenet that there are no relations of evidential

[2] This criterion adds the requirement that the negation of the evidence relevantly entails a local restriction of the negation of the hypothesis. Here, 'relevant entailment' is explicated by Ken Gemes's content part relation.

relevance between specific hypotheses and specific pieces of evidence – our scientific theories are instead revised by considerations of equilibrum affecting the entire field of research. I believe that it is possible to accept Duhem's point without giving in to Quine's contention: Admittedly, a piece of evidence does not only affect the hypothesis *under test*, but also the hypotheses *in use*. But generally, the degree of (dis)confirmation is quite different. Thus, the Duhem-Quine problem can be resolved in the framework of a quantitative theory of confirmation.

Quantitative accounts of confirmation usually explicate confirmation as increase in rational degree of belief in a hypothesis. To this end, chapter four has introduced a calculus for rational degrees of belief which are conceived as judgments about the fairness of hypothetical bets. Then it was demonstrated that this calculus conforms to the axioms of probability. This result, manifested in the Dutch Book theorems, is of twofold importance: On the one hand, it furnishes the calculus for degrees of belief with a tractable and well-developed mathematical theory, on the other hand, it forges a natural link between quantitative confirmation theory and statistical regularities which are expressed in the theory of probability. The interpretation of probabilities as rational degrees of beliefs constitutes the core of Bayesianism. Here, we have discussed the miscellaneous rationality constraints which the varieties of Bayesianism impose on rational credences. By the 'increase in rational credence' rationale, Bayesianism is naturally extended to a quantitative theory of confirmation which allows us to tackle the Duhem-Quine objection and to show that hypotheses under test and hypotheses in use can be supported to different degrees. However, we require *measures of support* to make that solution explicit, and this is the main topic of chapter five. A large set of measures of support that are well-known in the literature fails to satisfy mild adequacy criteria and falls through the grid. After consecutively narrowing down the list of admissible measures, the remaining measures explicate two quite different conceptions of confirmation and support: First, confirmation as a *generalization of logical entailment* and strength of an inductive argument, corresponding to a structural relationship between hypothesis and evidence. The log-likelihood measure l and the Crupi-Tentori measure z exemplify that conception. Second, confirmation as *impact* of the evidence on the epistemic status of the hypothesis and as actual increase in credibility. This is captured in the difference measure d. Both aspects are closely related to other virtues: posterior credibility of a hypothesis and informativeness of evidence. The second conception is, of course, highly contingent on the prior probability of the hypothesis which merely plays a subordinate role in the first explication. If there is at all a distinction between confirmation theory and inductive logic, we can locate it at this point: the analogy to deductive logic and the focus on the strength/validity of an inductive argument is typical of inductive logic whereas confirmation theory is more interested in the actual relevance of evidence for the tenability of a specific hypothesis. To my mind, it depends on the specific context of application which explication of inductive support should be preferred. To the best of my knowledge, this context-dependency of measures of support has not been elaborated in the existing literature. The context-dependency also transfers to the problem of old evidence where the way to resolve the problem depends on the concept of confirmation which one has in mind.

The most interesting and fruitful domain of application for quantitative confirmation theory consists in statistical regularities since statistical methods have recently made rapid progress in the empirical sciences (chapter six). Many physical and social processes are so complex that only the explicit incorporation of uncertainty and the use of probabilistic models are able to deliver adequate predictions of future events. Besides, the amount of available data has exponentially increased over the last decades, increasing the need for statistical analysis. Such a phenomenon is the more pronounced the more applied the field of inquiry. Since the application of statistical methods is nowadays pervasive in most empirical sciences, there is an increased demand for a thorough analysis of the foundations of statistical inference. Naturally, the Bayesian conception of inductive inference can be transferred to statistical inference, too. However, we recognize that mainstream Bayesianism, due to the subjective assignment of prior probabilities, cannot deliver a fully objective theory of inductive inference which is desired by many scientists who have to sell their results to peer-reviewed journals and fundraising agencies as ‘fully objective’ and ‘beyond the scope of sustainable subjective disagreement’. Of course, it is possible to ‘objectify’ Bayesian inference – the Maximum Entropy Principle constitutes the most prominent attempt. But the Maximum Entropy Principle represents uncertainty (as opposed to risk) and lack of information in specific ‘informationless’ prior distributions – an endeavor against which many criticisms can be raised (e.g. Seidenfeld 1979b, 1986). Be this as it may, although Bayesianism is by and large a subjective theory of inductive *inference*, its account of statistical *evidence* need not be equally subjective. Bayesian measures of statistical evidence as Bayes factors and (log-)likelihood ratios conform to Birnbaum's (1962) Likelihood Principle which asserts that all statistical evidence is just a function of the likelihood of the observed data under the competing hypotheses. Thus, subjective factors are ruled out on the whole.[3]

Although the Likelihood Principle is equivalent to the conjunction of two plausible elementary principles, most statistical practitioners (and philosophers of statistical inference) deliberately choose to violate it. Sociological factors put aside, the rationale for this decision consists in the conviction that valid inductive inference is characterized by the use of *reliable procedures*, i.e. procedures which generate correct conclusions in the vast majority of cases. This reliabilist approach to statistical inference is best articulated in *error statistics* (Mayo 1966, Mayo and Spanos 2006) – the lower the probability that a decision procedure errs, the more we should be confident to accept the result. The error-statistical and the Bayesian approach disagree on the question how to explicate statistical evidence, and the seventh chapter revolves around this question and connects it to the relevance of experimental design, i.e. the plans of an experimenter when to terminate a sequential trial.

The error-statistical concept of evidence is expressed in quantities as error probabilities, p-values, degrees of severity and significance levels. As measures of evidence, those quantities have several serious drawbacks, some of them located on a mathematical level whereas others are of conceptual nature. In particular, all error-statistical measures of evidence are sensitive to experimental design. This severely restricts scientific practice since those measures can be manipulated by insincerely reporting *intentions* about conduct and termination of an experiment without biasing the data. Therefore, those measures of evidence are not admissible which implies the evidential irrelevance of stopping rules and experimental design. Although the latter conclusions appear to be vulnerable and open to objections, a decision-theoretic perspective vindicates their soundness. However, it would be premature to infer from the evidential irrelevance of stopping rules, experimental design and error probabilities to their pre-experimental irrelevance – in fact, even for a Bayesian, error probabilities and experimental design affect the conduct of an experiment and the choice of a decision rule. I have merely defended

[3] Notably, this claim cannot be fully maintained for composite hypotheses because the relative weight of the single hypotheses has to be specified. But this is not only a problem for Bayesians, but also for non-Bayesians.

the claim that *after* observing the experimental results, those factors cannot play any inferential role – all that they could tell us is already contained in the observations.

Moreover, I have argued for a *comparative understanding of statistical evidence*. Thus, the task of finding a suitable measure of *support* which vexed us in chapter five has been replaced by the task of finding a suitable measure of *evidential favoring*, i.e. finding a quantity that measures to which degree observations favor a hypothesis over another one. It was already argued in chapter five that those questions are not coextensive. Bayesians often try to reduce evidential favoring to evidential support, but it has been argued that not all objections against specific measures of support (e.g. the log-ratio measure r) transfer to the induced measures of evidential favoring. On the contrary, a measure of evidential favoring that corresponds to a problematic measure of support (namely r) is the only one that satisfies a number of reasonable adequacy criteria (see Lele 2004 and the results in chapter seven). Adequacy criteria for measures of inductive support differ from criteria for measures of evidential favoring, corresponding to their different epistemic functions: While evidential support is especially important in situations where few or no definite competitors to the hypothesis under test can be identified, evidential favoring becomes central whenever there are various or many competing hypotheses – a situation characteristic of statistical reasoning. Therefore it is not surprising that statistical confirmation theory focuses on evidential favoring (chapter seven) while non-statistical confirmation theory (chapter two, three and five) focuses on evidential relevance and support.

In the remainder, I would like to sketch some more general conclusions and mention open questions for future research. First, the foundational superiority of statistical inference that complies with the Likelihood Principle over the error-statistical competitors suggests that the search for a universal and objective account of inductive inference that gives a sound explication of statistical evidence may be in vain. Subjective elements are required to structure complex composite hypotheses and to assign relative weights to the single elements of such a composite hypothesis. Such a subjective approach might look undesirable since the pursuit of absolute objectivity is abandoned. But first, assessing composite hypotheses without knowing the (subjectively fixed) relative weights of the single parts is just impossible. For instance, the less credible we find a hypothesis a priori, the less relative weight should it have in a composite hypothesis to which it belongs. Second, subjective assignments of prior probabilities can express scientific expertise, too, and ultimately, it is only logical that the premises of an inductive inference – the prior opinions – affect the result, too. Thus, we should no longer discard Bayesian inference on the grounds of its subjectivity – first, there is no sensible alternative and second, subjectivity need not be a vice. Clearly, the more complex matters get and the more competing hypotheses are involved, the less can qualitative accounts do the job, and even the falsificationist criterion comes to its limits, due to its lack of a quantitative dimension. Those criteria are valuable for modeling confirmation in the history of science, but they fail to be applicable to modern statistical reasoning.

Second, it is interesting to transfer the results on Bayesian and error statistics to the thesis of epistemic reliabilism: Agents that entertain belief X are justified if and only if that belief was generated by a reliable causal process, i.e. a process that tends to produce much more true than false beliefs (see Goldman 1979). In a similar vein, error statistical methods as confidence intervals justify conclusions by the reliability of the belief-forming procedure (here: the method used to construct the confidence interval). It is, however, unclear whether such a justification really helps us since we are not interested in the reliability of statistical methods that lead to a result R (e.g. a specific confidence interval) but in the *actual correctness* of R and the question whether we can put *confidence* in R and base our decisions on the belief that R is true (see the Pratt quote on page 174). Rational decisions are – this is another result of statistical decision theory – solely based on posterior probabilities. Mathematically, it is impossible to infer from the reliability of a method that yields the result R to a rational posterior credence in R. This insight falls back on epistemic reliabilists as long as they want to maintain a link between epistemic justification and rational decisions. Two related issues that are suitable for further research deserve mention: First, the epistemological debate fails to consider and to examine the contrast between counterfactual reasoning (which is exemplified in reliabilism) and reasoning that is merely based on actual observations (as stated in the Conditionality Principle). In particular, it is exciting to see whether the arguments that were given for inductive reasoning according to the Likelihood and Conditionality Principle directly count against epistemic reliabilism.

Third, statistics, the rapidly developing science of inductive reasoning, deserves more attention from philosophical perspective. For a long time, statistics has been neglected by philosophers, Teddy Seidenfeld and Deborah Mayo being notable exceptions. It seemed to be either a part of mathematics or an appendix to the natural sciences, and the special role of statistics between the poles of inductive practice (science) and foundational rigor (mathematics) has not been acknowledged from a philosophical perspective. Most of the literature on the statistical foundations of inductive inference has been written by statisticians, among them people like James O. Berger or Richard Royall who show a surprising openness to philosophical and methodological questions (actually, some issues are sometimes explicitly described as 'philosophical'). Time is ripe that these issues are resumed by philosophers in order to initiate an intense and fruitful exchange with statistical practitioners. In particular, the confusion about p-values and other statistical methods in the empirical sciences clearly asks for a methodological clarification which in turn calls for philosophical expertise. Moreover, the foundational work on the principles of inductive inference that has been done in statistics (e.g. Birnbaum's derivation of the Likelihood Principle) has not yet been fully recognized in the philosophical community. Finally, the topic of chapter seven, the inferential relevance of experimental design, gains social relevance in the context of evidence-based decisions. Politicians often base their decisions on competent scientific advice (at least they are supposed to do so), and those advisors should be able to discern which factors (e.g. experimental design) are relevant to the evidence which an experiment yields.

For all these reasons, I believe that the most interesting and exciting open questions about inductive reasoning are found in statistics and not in classical confirmation theory or inductive logic although the latter can certainly give fruitful impulses. Among those open questions are, for example: How can the methodology of model analysis and selection be characterized from a philosophical perspective? In particular, what is the relationship between simplicity, goodness-of-fit and overall adequacy of a model? What is the function of relative unexpectedness and surprise in a model selection analysis? How do statisticians deal with the old evidence problem? Another major and almost completely neglected issue is the inferential role of nonparametric reasoning – statistical techniques that do not take specific (parametric) families of distributions as given and instead rely on 'qualitative' constraints (such as the expectation value and the variance of the unknown distribution). This book stayed in the realms of parametric statistical models, partly for reasons of space, but nonparametric reasoning is actually a fast-growing and important kind of statistical inference which has to be examined in future work.

A lot of question has been left open and some exciting issues could not even be discussed in the relative brevity of this book. Other problems could have been scrutinized in more depth. Nevertheless I hope that the reader has gained some illuminating insights into the principles of inductive inference and realized the central position of this field of research in philosophy of science and beyond.

Appendix A

Proof of the Dutch Book theorem

We sketch the proof of the Dutch Book theorem here – readers without interest in the mathematics can skip it, those with interest in the details may have a look at Kemeny (1955) or Skyrms (1980). We start with the first axiom. If any tautology were assigned a degree of belief less than 1, the associated betting odds would be greater than 1 (because probability p is mapped to fair betting odds $1/p$). But this would allow to bet on a tautology and to get back more money than the stake. This offers a riskless gain to the bettor and is therefore no fair bet.

Now we come to the second axiom. Assume first that $P(A) + P(\neg A) < 1$ and that neither A nor $\neg A$ is a tautology. Then, by a series of equivalent transformations, it follows that

$$\left(\frac{1}{P(\neg A)} - 1\right)^{-1} < \frac{1}{P(A)} - 1 \tag{A.1}$$

Now, choose a y so that

$$\frac{1}{\frac{1}{P(\neg A)} - 1} < y < \frac{1}{P(A)} - 1 \tag{A.2}$$

Such a y exists because of (A.1). Now, we propose a following system of fair bets corresponding to the probabilities for A and $\neg A$ (table A.1):

In other words, we bet on A with stake 1 Euro and on $\neg A$ with stake y Euro, where the return is given by the probabilities on A and $\neg A$ and the associated betting odds. Either A or $\neg A$ has to occur, by the law of

$\langle A \mid 1 \mid 1/P(A) - 1\rangle$
$\langle \neg A \mid y \mid y/P(\neg A) - 1\rangle$

Table A.1: A system of bets.

$\langle A_1 \vert P(A_1) \vert 1 - P(A_1)\rangle$
$\langle A_2 \vert P(A_2) \vert 1 - P(A_2)\rangle$
$\langle A_3 \vert P(A_3) \vert 1 - P(A_3)\rangle$
$\ldots$
$\big\langle \neg(\vee_{n\in\mathbb{N}} A_n) \mid P(\neg \bigvee_{n\in\mathbb{N}} A_n) \mid 1 - P(\neg \bigvee_{n\in\mathbb{N}} A_n)\big\rangle$

Table A.2: Another system of bets.

the excluded middle. Regardless of whether A or $\neg A$ occurs, the bettor is guaranteed a positive net return. Indeed, if A occurs, the bettor receives

$$\frac{1}{P(A)} - 1 - y > 0 \tag{A.3}$$

due to the second inequality in (A.2). If $\neg A$ occurs, the bettor receives

$$\frac{y}{P(\neg A)} - 1 - y = y\left(\frac{1}{P(\neg A)} - 1\right) - 1 > 0 \tag{A.4}$$

due to the first inequality in (A.2). Hence, equations (A.3) and (A.4) show that the proposed system of bets promises a safe win for the bettor if he bets according to the 'fair' odds. But then, the game is no longer a zero-sum game and the odds have not been fair. A similar argument can be made if $P(A) + P(\neg A) > 1$. Hence, the second axiom of probability is mandatory for degrees of belief if Dutch Books are to be avoided.

Finally, the third axiom. Consider a series of mutually exclusive events $A_1, A_2, A_3, \ldots$. Now we construct the following system of individually fair bets (table A.2). The system is so designed that exactly one of the bets is going to win in any case. Hence, the net return R of the bettor is independent of the course of events and identical to

$$\begin{aligned} R &= 1 - \sum_{n\in\mathbb{N}} P(A_n) - P(\neg \vee_{n\in\mathbb{N}} A_n) \\ &= 1 - \sum_{n\in\mathbb{N}} P(A_n) - 1 + P(\vee_{n\in\mathbb{N}} A_n) \\ &= P(\vee_{n\in\mathbb{N}} A_n) - \sum_{n\in\mathbb{N}} P(A_n) \end{aligned}$$

Now, assume that $P(\bigvee_{n\in\mathbb{N}} A_n) > \sum_{n\in\mathbb{N}} P(A_n)$, in violation of the third Kolmogorov axiom. Then the system of bets in table A.2 cannot have been fair since it guarantees the bettor a positive net gain. Hence, a Dutch Book has been construed. Similarly, if $P(\bigvee_{n\in\mathbb{N}} A_n) < \sum_{n\in\mathbb{N}} P(A_n)$, the above system will yield a positive net gain for the bookie, with equally devastating consequences. Violating the third axiom of probability allows the construction of Dutch Books. Hence, all three axioms of probability have to be obeyed when Dutch books shall be avoided. $\square$

Appendix B

Sequential trials: details for table 7.2

In a sequential trial, the experimental costs have to be deduced from the total utilities. If costs are zero, the utility matrices are the same for each experimental design. Assume now that each sample entails fixed costs. In a fixed sample size experiment with $N = 12$, the same number is deduced from each element of the matrix, see the middle columns of table 7.2. Finally, sampling might terminate after a fixed number of successes K. Then, the expected number of samples $N(H_i)$ under the competing hypotheses $H_0 : p = 1/2$ and $H_1 : p = 1/4$ is

$$\begin{aligned} N(H_0) &= K + K\frac{1 - p_{H_0}}{p_{H_0}} = 2K = 6 \\ N(H_1) &= K + K\frac{1 - p_{H_1}}{p_{H_1}} = 4K = 12 \end{aligned}$$

This explains the numbers in the right part of table 7.2.

Bibliography

Armitage, Peter (1975): *Sequential Medical Trials.* Oxford, Blackwell.

Armitage, Peter and Geoffrey Berry (1987): *Statistical Methods in Medical Research.* Second Edition. Springer, New York.

Barnard, George A. (1949): "Statistical Inference (with Discussion)", *Journal of the Royal Statistical Society, series B* 11, 115-139.

Bayarri, M. J. and James O. Berger (2000): "P Values for Composite Null Models", *Journal of the American Statistical Association* 95, 1127-1142.

Berger, James O. (1985): *Statistical Decision Theory and Bayesian Analysis.* Second Edition. Springer, New York.

Berger, James O. and Donald Berry (1988): "The Relevance of Stopping Rules in Statistical Inference (with discussion)", in: S. Gupta and J. O. Berger (ed.), *Statistical Decision Theory and Related Topics IV*, 29-72. Springer, New York.

Berger, James O. and Thomas Sellke (1987): "Testing a Point Null Hypothesis: the Irreconcilability of P Values and Evidence", *Journal of the American Statistical Association* 82, 106-111.

Berger, James O. and Robert L. Wolpert (1984): *The Likelihood Principle.* Institute of Mathematical Statistics, Hayward/CA.

Bernardo, J. M. (1979): "Reference posterior distributions for Bayesian inference", *Journal of the Royal Statistical Society, series B* 41, 113-147.

Birnbaum, Allan (1962): "On the Foundations of Statistical Inference", *Journal of the American Statistical Association* 57, 269-306.

Birnbaum, Allan (1972): "More on Concepts of Statistical Evidence", *Journal of the American Statistical Association* 67, 858-861.

Carnap, Rudolf (1950): *Logical Foundations of Probability.* The University of Chicago Press, Chicago.

Casella, George and Roger L. Berger (1987): "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem", *Journal of the American Statistical Association* 82, 106-111.

Christensen, David (1983): "Glymour on Evidential Relevance", *Philosophy of Science* 50, 471-481.

Christensen, David (1990): "The Irrelevance of Bootstrapping", *Philosophy of Science* 57, 644-662.

Christensen, David (1999): "Measuring Confirmation", *Journal of Philosophy* 96, 437-461.

Colyvan, Mark (2001): *The Indispensability of Mathematics.* Oxford University Press, Oxford.

Crupi, Vincenzo, Katya Tentori and Michel Gonzalez (2007): "On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues", *Philosophy of Science* 74, 229-252.

Crupi, Vincenzo, Branden Fitelson and Katya Tentori (2008): "Probability, Confirmation, and the 'Conjunction Fallacy'", forthcoming in *Thinking and Reasoning.*

De Finetti, Bruno (1937): "La prévision: Ses lois logiques, ses sources objectives", *Annales de l'Institut Henri Poincaré* 7, 1-68.

Dennis, Brian (2004): "Statistics and the Scientific Method in Ecology (with discussion)", in: Mark Taper and Subhash Lele (ed.), *The Nature of Scientific Evidence*, 327-378. The University of Chicago Press, Chicago & London.

Douven, Igor and Wouter Meijs (2006): "Bootstrap Confirmation Made Quantitative", *Synthese* 149, 97-132.

Duhem, Pierre (1914): *La Théorie Physique: Son Objet, Sa Structure.* Second edition, reprinted in 1981 by J. Vrin, Paris.

Earman, John (1992): *Bayes or Bust?.* The MIT Press, Cambridge/MA.

Earman, John and Clark Glymour (1992): "The Confirmation of Scientific Hypotheses", in: Merrilee H. Salmon (ed.), *Introduction to the Philosophy of Science*, 42-103. Hackett, Indianapolis.

Edwards, Ward, Harold Lindman and Leonard J. Savage (1963): "Bayesian Statistical Inference for Psychological Research", *Psychological Review* 70, 450-499.

Edwards, A. W. F. (1992): *Likelihood.* Second Edition, Cambridge University Press, Cambridge.

Eells, Ellery and Branden Fitelson (2002): "Symmetries and Asymmetries in Evidential Support", *Philosophical Studies* 107, 109-122.

Festa, Roberto (1986): "A measure for the distance between an interval hypothesis and the truth", *Synthese* 67, 273-320.

Festa, Roberto (1993): *Optimum Inductive Methods.* Kluwer, Dordrecht.

Fisher, Ronald A. (1959): *Statistical Methods and Scientific Inference.* Second Edition, Hafner, New York.

Fitelson, Branden (1999): "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity", *Philosophy of Science* 66, S362-S378.

Fitelson, Branden (2001a): *Studies in Bayesian Confirmation Theory.* PhD thesis, University of Wisconsin, Madison.

Fitelson, Branden (2001b): "A Bayesian Account of Independent Evidence with Applications", *Philosophy of Science* 68, S123-S140.

Fitelson, Branden (2002): "Putting the Irrelevance Back into the Problem of Irrelevant Conjunction", *Philosophy of Science* 69, 611-622.

Fitelson, Branden (2005): "Inductive Logic", in: J. Pfeifer, S. Sarkar (ed.), *Philosophy of Science: An Encyclopedia.* Routledge, London.

Fitelson, Branden (2006): “The Paradox of Confirmation”, *Philosophy Compass* 1, 95–113.

Fitelson, Branden (2007): “Likelihoodism, Bayesianism and Relational Confirmation”, *Synthese* 156, 473-489.

Fitelson, Branden (2008): “Goodman’s ‘New Riddle’”, forthcoming in *Journal of Philosophical Logic.*

Fitelson, Branden and James Hawthorne (2006): “How Bayesian Confirmation Theory Handles the Paradox of the Ravens”, in: Ellery Eells and James Fetzer (ed.), *Probability in Science.* Open Court, Chicago.

Fitelson, Branden and Andrew Waterman (2005): “Bayesian Confirmation and Auxiliary Hypotheses Revisited: A Reply to Strevens”, *British Journal for the Philosophy of Science* 56, 293-302.

Fitelson, Branden and Andrew Waterman (2007): “Comparative Bayesian Confirmation and the Quine-Duhem Problem: A Rejoinder to Strevens”, *British Journal for the Philosophy of Science* 58, 333-338.

Gaifman, Haim (1979): “Subjective Probability, Natural Predicates and Hempel’s Ravens”, *Erkenntnis* 21, 105-147.

Gaifman, Haim and Marc Snir (1982): “Probabilities over Rich Languages”, *Journal of Symbolic Logic* 47, 495–548.

Garber, Dan (1983): “Old Evidence and Logical Omniscience in Bayesian Confirmation Theory”, in: John Earman (ed.), *Testing Scientific Theories. Minnesota Studies in the Philosophy of Science, vol. 10.* University of Minnesota Press, Minnesota.

Gemes, Ken (1993): “Hypothetico-Deductivism, Content and the Natural Axiomatisation of Theories”, *Philosophy of Science* 60, 477-487.

Gemes, Ken (1994): “A New Theory of Content”, *Journal of Philosophical Logic* 23, 596-620.

Gemes, Ken (1997): “A New Theory of Content II: Model Theory and Some Alternatives”, *Journal of Philosophical Logic* 26, 449-476.

Gemes, Ken (1998): "Hypothetico-Deductivism: The Current State of Play", *Erkenntnis* 49, 1-20.

Gemes, Ken (2006a): "Content and Watkins' Account of Natural Axiomatizations", *dialectica* 60, 85-92.

Gemes, Ken (2006b): "Bootstrapping and Content Parts", *Erkenntnis* 64, 345-370.

Georgii, Hans-Otto (2002): *Stochastik.* De Gruyter, Berlin.

Gillies, Donald (1986): "In Defense of the Popper-Miller Argument", *Philosophy of Science* 53, 110-113.

Glymour, Clark (1975): "Relevant Evidence", *Journal of Philosophy* 72, 403-426.

Glymour, Clark (1980a): *Theory and Evidence.* Princeton University Press, Princeton.

Glymour, Clark (1980b): "Discussion: Hypothetico-Deductivism is Hopeless", *Philosophy of Science* 47, 322-325.

Glymour, Clark (1983): "Revisions of Bootstrap Testing", *Philosophy of Science* 50, 626-629.

Goldman, Alvin (1979): "What is justified belief?", in: G. Pappas and M. Swain (ed.), *Justification and Knowledge*, 1-23. Reidel, Dordrecht.

Good, I. J. (1967): "The White Shoe is a Red Herring", *British Journal for the Philosophy of Science* 17, 322.

Good, I. J. (1968): "The White Shoe qua Herring is Pink", *British Journal for the Philosophy of Science* 19, 156-157.

Good, I. J. (1983): *Good Thinking: The Foundations of Probability and Applications.* University of Minnesota Press, Minneapolis.

Goodman, Nelson (1983): *Fact, Fiction and Forecast.* Fourth Edition. Harvard University Press, Cambridge/MA.

Grimes, Thomas R. (1990): "Truth, Content, and the Hypothetico-Deductive Method", *Philosophy of Science* 57, 514-522.

Hacking, Ian (1965): *Logic of Statistical Inference.* Cambridge University Press, Cambridge.

Haenni, Rolf, Jan-Willem Romeijn, Gregory Wheeler and Jon Williamson (2008): "Possible Semantics for a Common Framework of Probabilistic Logics", in: V. N. Huynh (ed.), *Proceedings of the International Workshop on Interval/Probabilistic Uncertainty and Non-Classical Logics.* Advances in Soft Computing, Ishikawa.

Hailperin, Theodore (1984): "Probability Logic", *Notre Dame Journal of Formal Logic* 25, 198-212.

Hailperin, Theodore (1996): *Sentential Probability Logic: Origins, Development, Current Status and Technical Applications.* Lehigh University Press, Bethlehem/PA.

Hájek, Alan (2003): "What Conditional Probability Could Not Be", *Synthese* 137, 273-323.

Hájek, Alan (2007): "Interpretations of Probability", in: Stanford Encyclopedia of Philosophy, retrieved at: *http://plato.stanford.edu/entries/probability-interpret/*

Hawthorne, James (2008): "Inductive Logic", in: Stanford Encyclopedia of Philosophy, retrieved at: *http://plato.stanford.edu/entries/logic-inductive*

Hawthorne, James and Branden Fitelson (2004): "Re-solving Irrelevant Conjunction with Probabilistic Independence", *Philosophy of Science* 71, 505-514.

Hempel, Carl G. (1943): "A Purely Syntactical Definition of Confirmation", *Journal of Symbolic Logic* 8, 122-143.

Hempel, Carl G. (1965): "Studies in the Logic of Confirmation", in: *Aspects of Scientific Explanation*, 3-46. The Free Press, New York. Reprint from *Mind* **54**, 1945.

Horwich, Paul (1982): *Probability and Evidence.* Cambridge University Press, Cambridge.

Howson, Colin (2003): *Hume's problem.* Oxford University Press, Oxford.

Howson, Colin and Peter Urbach (1993): *Scientific Reasoning: The Bayesian Approach.* Second Edition. Open Court, La Salle.

Huber, Franz (2005): "What Is the Point of Confirmation?", *Philosophy of Science* 72, 1146-1159.

Hume, David (1777): *An Enquiry Concerning Human Understanding.* Posthumous edition, reprinted in 1902 by Clarendon Press, Oxford (edited by L. A. Selby-Bigge).

Jaynes, Edwin T. (1957): "Information Theory and Statistical Mechanics I+II", *Physical Review* 106+108, 620-630, 171-190.

Jaynes, Edwin T. (1968): "Prior Probabilities", *I.E.E.E. Transactions on System Science and Cybernetics* SSC-4, 227-41.

Jeffrey, Richard C. (1983): "Bayesianism with a Human Face", in: John Earman (ed.), *Testing Scientific Theories. Minnesota Studies in the Philosophy of Science, vol. 10.* University of Minnesota Press, Minnesota.

Jeffreys, Harold (1961): *Theory of Probabiliy.* Third Edition, Oxford University Press, Oxford.

Joyce, James (1999): *The Foundations of Causal Decision Theory.* Cambridge University Press, Cambridge.

Joyce, James (2003): "Bayes's Theorem", in: The Stanford Encyclopedia of Philosophy, retrieved at: *http://plato.stanford.edu/archives/win2003/entries/bayes-theorem/*

Kadane, Joseph B., Mark J. Schervish and Teddy Seidenfeld (1996a): "Reasoning to a Foregone Conclusion", *Journal of the American Statistical Association* 91, 1228-1236.

Kadane, Joseph B., Mark J. Schervish and Teddy Seidenfeld (1996b): "When Several Bayesians Agree That There Will Be No Reasoning to a Foregone Conclusion", *Philosophy of Science* 63, S281-S289.

Kemeny, John G. (1955): "Fair Bets and Inductive Probabilities", *Journal of Symbolic Logic* 20, 263-273.

Kemeny, John G. and Paul Oppenheim (1952): "Degrees of Factual Support", *Philosophy of Science* 19, 307-324.

Keynes, John M. (1921): *A Treatise on Probability.* Macmillan, London.

Kolmogorov, Andrey Nikolaevich (1956): *Foundations of the Theory of Probability.* Original work published in German in 1933. Chelsea, New York.

Kuhn, Thomas S. (1962): *The Structure of Scientific Revolutions.* The University of Chicago Press, Chicago.

Kullback, S. and R. A. Leibler (1951): "On Information and Sufficiency", *Annals of Mathematical Statistics* 22, 79-86.

Laplace, Pierre Simon (1951): *A Philosophical Essay on Probabilities.* Original work published in French in 1814. Dover, New York.

Lele, Subhash (2004): "Evidence Functions and the Optimality of the Law of Likelihood (with discussion)", in: Mark Taper and Subhash Lele (ed.), *The Nature of Scientific Evidence*, 191-216. The University of Chicago Press, Chicago & London.

Lewis, David (1980): "A Subjectivist's Guide to Objective Chance", in: Richard C. Jeffrey (ed.), *Studies in Inductive Logic and Probability.* University of California Press, Berkeley.

Lindley, Dennis (1957): "A Statistical Paradox", *Biometrika* 44, 187-192.

Lindley, Dennis (2000): "The Philosophy of Statistics", *Journal of the Royal Statistical Society, series D* 49, 293-337.

Maher, Patrick (2006): "The Concept of Inductive Probability", *Erkenntnis* 65, 185-206.

Mayo, Deborah G. (1996): *Error and the Growth of Experimental Knowledge.* The University of Chicago Press, Chicago & London.

Mayo, Deborah G. and Michael Kruse (2001): "Principles of inference and their consequences", in: D. Cornfield, J. Williamson (ed.), *Foundations of Bayesianism*, 381-403. Kluwer, Dordrecht.

Mayo, Deborah G. and Aris Spanos (2006): “Severe Testing as a Basic Concept in a Neyman-Person Philosophy of Induction”, *British Journal for the Philosophy of Science* 57, 323-357.

Meijs, Wouter (2005): *Probabilistic Measures of Coherence.* PhD thesis, Erasmus University Rotterdam.

Mellor, Hugh (2005): *Probability: A Philosophical Introduction.* Routledge, London.

Milne, Peter (1996): “log[p(h/eb)/p(h/b)] is the one true measure of confirmation”, *Philosophy of Science* 63, 21-26.

Moretti, Luca (2004): “Grimes on the Tacking by Disjunction Problem”, *disputatio* 1, 16-20.

Moretti, Luca (2006): “The Tacking by Disjunction Paradox: Bayesianism versus Hypothetico-Deductivism”, *Erkenntnis* 64, 115-138.

Morrsion, Joe (2008): *Just How Controversial is Evidential Holism?*, forthcoming manuscript.

Mortimer, Halina (1988): *The Logic of Induction.* Halsted Press, New York.

Nicod, Jean (1970): *Geometry and Induction.* University of California Press, Berkeley and Los Angeles. English translation of works originally published in French in 1923 and 1924.

Niiniluoto, Ilkka (1983): “Novel Facts and Bayesianism”, *British Journal for the Philosophy of Science* 34, 375-379.

Nozick, Robert (1981): *Philosophical Explanations.* Clarendon Press, Oxford.

Popper, Karl R. (1963): *Conjectures and Refutations: The Growth of Scientific Knowledge.* Routledge, London.

Pratt, J. W. (1961): “Review of Lehmann, E. L.: ‘Testing Statistical Hypotheses’”, *Journal of the American Statistical Association* 56, 163-167.

Quine, Willard Van Orman (1961): “Two Dogmas of Empiricism”, in: *From a Logical Point of View*, 20-46. Second Edition, Harvard University Press, Cambridge/MA.

Ramsey, Frank P. (1978): "Truth and Probability", in: Hugh Mellor (ed.), *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*, 58-100. Routledge, London. Original article published in 1926..

Rawls, John (1971): *A Theory of Justice*. Revised Edition. Harvard University Press, Cambridge/MA.

Reichenbach, Hans (1956): *The Direction of Time*. University of California Press, Berkeley.

Rips, Lance J. (2001): "Two Kinds of Reasoning", *Psychological Science* 12, 129-134.

Robins, James M., Aad van der Vaart and Valérie Ventura (2000): "The asymptotic distribution of p-values in composite null models", *Journal of the American Statistical Association* 95, 1143-1156.

Rosenkrantz, R. D. (1981): *Foundations and Applications of Inductive Probability*. Ridgeview, Atascadero.

Rosenkrantz, R. D. (1994): "Bayesian Confirmation: Paradise Regained", *British Journal for the Philosophy of Science* 45, 467-476.

Rosenthal, Jacob (2004): *Wahrscheinlichkeiten als Tendenzen*. Mentis, Paderborn.

Royall, Richard (1997): *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London.

Royall, Richard (2000): "On the Probability of Observing Misleading Statistical Evidence", *Journal of the American Statistical Association* 95, 760-768.

Schurz, Gerhard (1991): "Relevant Deduction", *Erkenntnis* 35, 391-437.

Schurz, Gerhard (1994): "Relevant Deduction and Hypothetico-Deductivism: A Reply to Gemes", *Erkenntnis* 41, 183-188.

Schurz, Gerhard (2005): "Bayesian H-D Confirmation and Structuralistic Truthlikeness: Discussion and Comparison with the Relevant-Element and the Content-Part Approach", in: Roberto Festa (ed.), *Logics of Scientific*

Discovery. Essays in Debate with Theo Kuipers, 141-159. Rodopi, Amsterdam.

Seidenfeld, Teddy (1979a): *Philosophical Problems of Statistical Inference: Learning from R. A. Fisher*. Reidel, Dordrecht.

Seidenfeld, Teddy (1979b): "Why I am not an Objective Bayesian", *Theory and Decision* 11, 413-440.

Seidenfeld, Teddy (1986): "Entropy and Uncertainty", *Philosophy of Science* 53, 467-491.

Sellke, Thomas, M. J. Bayarri and James O. Berger (2001): "Calibration of P-Values for testing precise null hypotheses", *The American Statistician* 55, 62-71.

Selvin, Hanan (1970): "A critique of tests of significance in survey research", in: D. Morrison, R. Henkel (ed.), *The significance test controversy*, 94-106. Aldine, Chicago.

Shannon, Claude and Warren Weaver (1949): *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Sober, Elliott (1999): "Testability", *Proceedings and Addresses of the American Philosophical Association* 73, 47-76.

Spohn, Wolfgang (1990): "A General Non-Probabilistic Theory of Inductive Reasoning", in: R.D. Shachter, T.S. Levitt, J. Lemmer, L.N. Kanal (ed.), *Uncertainty in Artificial Intelligence 4*, 149-158. Elsevier, Amsterdam.

Spohn, Wolfgang (2008): "A Survey of Ranking Theory", forthcoming in Franz Huber, Christoph Schmidt-Petri (ed.), *Degrees of Belief. An Anthology*. Oxford University Press, Oxford.

Staley, Kent (2004): "Robust Evidence and Secure Evidence Claims", *Philosophy of Science* 71, 467–88.

Strevens, Michael (2001): "The Bayesian Treatment of Auxiliary Hypotheses", *British Journal for the Philosophy of Science* 52, 515-537.

Strevens, Michael (2005): "The Bayesian Treatment of Auxiliary Hypotheses: Reply to Fitelson and Waterman", *British Journal for the Philosophy of Science* 56, 913-918.

Taper, Mark and Subhash Lele (eds.) (2004): *The Nature of Scientific Evidence.* The University of Chicago Press, Chicago & London.

Tentori, Katya, Vincenzo Crupi, Nicolao Bonini and Daniel Osherson (2007): "Comparison of Confirmation Measures", *Cognition* 103, 107-119.

Uffink, Jos (1996): "The Constraint Rule of the Maximum Entropy Principle", *Studies in the History and Philosophy of Modern Physics* 96, 47-79.

Wald, Abraham (1947): *Sequential Analysis.* John Wiley, New York.

Wald, Abraham (1950): *Statistical Decision Functions.* John Wiley, New York.

Wiechert, Emil (1920): "Die Gravitation als elektrodynamische Erscheinung", *Annalen der Physik* 63, 301-381.

Williams, P. M. (1980): "Bayesian Conditionalisation and the Principle of Minimum Information", *British Journal for the Philosophy of Science* 31, 131-144.

Williamson, Jon (2005): "Objective Bayesian Nets", in: S. Artemov, H. Barringer, A.S. d'Avila Garcez, L.C. Lamb and J. Woods (ed.), *We Will Show Them: Essays in Honour of Dov Gabbay, Vol. 2*, 713-730. College Publications, London.

Williamson, Jon (2007): "Motivating objective Bayesianism: from empirical constraints to objective probabilities", in: William Harper, Gregory Wheeler (ed.), *Probability and Inference: Essays in Honour of Henry E. Kyburg Jr.*, 151-179. College Publications, London.

List of Figures

List of Tables