29

Empirical Evidence: Its Nature and Sources

Julian Reiss

INTRODUCTION

With the rise of evidence-based movements in medicine and social policy, the topic of evidence has come to the forefront of research in the philosophy and methodology of science. But the issue is far from new. Observation, experiment, induction and confirmation – all practises very closely related to evidence – have been central concerns of philosophers ever since the birth of Western philosophy. The primary aim of this article is to provide an introduction to and illumination of these topics in so far as they are relevant to the social sciences.

Empirical evidence in the social sciences is extraordinarily varied. It is produced by methods including the collection of physical artefacts in archaeology, conducting censuses in demography, mathematical modelling in economics, thought experimentation in history, expert judgement in political science, laboratory experimentation in psychology and causal modelling in sociology, among many others. Even within one and the same science, evidence can have a variety of sources. To take economic policy as an example, the traditional way to substantiate economic policy claims is to first build a structural model, which is based on economic theory, then to operationalise its terms and test the fit of the model against data. This structural equation modelling approach was developed by members of the Cowles Commission in the 1930s and 1940s and has since been supplemented by a variety of other techniques in econometrics such as the analysis of natural experiments, but also laboratory experiments, simulation and various forms of conceptual or thought experimentation.

This article has two parts. In the first part, I will survey philosophical theories of evidence and in so doing attempt to answer questions regarding the nature of evidence and the nature of the inference from evidence to hypotheses. The second part will give a necessarily incomplete overview of the different sources of evidence in the social sciences and asks how to combine their products.

Before diving into the various philosophical proposals for theories of evidence and inference, a few terminological clarifications are in order. First, the notion of scientific evidence has at least three connotations that are relevant here. According to *Webster's New World Dictionary* (Second College edition), 'evidence' refers to: (1) the condition

5579-Jarvie-Chap29.indd 551 11/10/2010 3:26:05 PM

of being evident; (2) something that makes another thing evident; indication; sign; and (3) something that tends to prove; ground for belief. Accordingly, *scientific* evidence means, first, the more or less observable outcomes of scientific tests such as experiments, statistical analyses and surveys. Used in this way, the notion is more or less synonymous with 'data' or 'scientific result'. According to the second entry, scientific evidence means hint, sign, indication of or a reason to believe (the negation of) a scientific hypothesis. According to the third, the word means (something that furnishes) proof of or good or cogent reason to believe (the negation of) a hypothesis. The ambiguity between the latter two meanings is illustrated by a passage from Wesley Salmon (Salmon, 1975: 6):

> As Carnap pointed out in *Logical Foundations of Probability*, the concept of confirmation is radically ambiguous. If we say, for example, that the special theory of relativity has been confirmed by experimental evidence, we may have either of two quite distinct meanings in mind. On the one hand, we may intend to say that the special theory has become an accepted part of scientific knowledge and that it is very nearly certain in light of its supporting evidence. If we admit that scientific hypotheses can have numerical degrees of confirmation, the sentence, on this construal, says that the degree of confirmation of the special theory of relativity on the available evidence is high. On the other hand, the same sentence might be used to make a very different statement. It might be taken to mean that some particular evidence – for example, observations on the lifetimes of mesons – renders the special theory more acceptable or better founded than it was in the absence of this evidence. If numerical degrees of confirmation are again admitted, this latter construal of the sentence amounts to the claim that the special theory has a higher degree of confirmation on the basis of the new evidence than it had on the basis of the previous evidence alone.

In what follows I will use 'evidence' almost always in sense 2, indication or sign, the only exception being Peter Achinstein's theory of evidence, which is a theory of a good or cogent reason to believe. Evidence in sense 1 will play an important role but I will say 'data' or 'test result' or something similar for it.

Second, as the passage from Salmon illustrates, the notion of evidence is closely related to those of confirmation and induction. Hypotheses are confirmed by evidence, and most theories of evidence that will be discussed below have been introduced as theories of confirmation. Induction refers to the mode of reasoning or inference from evidence to hypothesis. As the hypothesis contains more, or at least different, information than the evidence, this mode of reasoning is ampliative – enlarging what is already known. It contrasts with deductive reasoning, which proceeds in reverse order from a more general claim to a more specific claim. Unlike deductive reasoning, which is truth preserving, reasoning from evidence is fallible.

TAKING THE CON OUT OF CONFIRMATION

Some Preliminary Remarks

Philosophers of science often treat theories of evidence and of induction (or confirmation) as if they were of the same kind. But this would be a mistake. Theories of evidence are supposed to answer questions regarding the nature of evidence and the kinds of observations or tests a researcher needs to make in order to have evidence in favour of the hypothesis at stake. Theories of induction, by contrast, begin with an antecedently understood notion of evidence and ask what the kinds of inferences are that one is justified to make given one has evidence of the required type at hand.

Bayesianism, for example, is often described as a theory of evidence (as in Achinstein, 2001) but it is completely silent on the issue of the nature of evidence. Rather, it tells us what a rational agent should do in a situation where she comes to believe an evidential statement (*viz*., to update her

degree of belief in the hypothesis according to a specific rule). Colin Howson and Peter Urbach describe the matter as follows:

> The Bayesian theory we are proposing is a theory of inference from data; we say nothing about whether it is correct to accept the data … The Bayesian theory of support is a theory of how the acceptance as true of some evidential statement affects your belief in some hypothesis. How you came to accept the truth of the evidence, and whether you are correct in accepting it as true, are matter which, from the point of view of the theory, are simply irrelevant. (Howson and Urbach, 1993: 272)

By contrast, Mill's methods, though often described as a theory of induction (as in Norton, 2003) are in fact at best informative about the types of observations one must make in order to support a (in this case, causal) hypothesis. For instance, the method of difference tells us evidence in favour of a causal hypothesis can be provided by two situations that are exactly identical except with respect to the phenomenon of interest. The method, by itself, is not informative about the kinds of inferences warranted by the observation of two such situations. It does not say, for example, whether, after having made the observation, we should accept the causal hypothesis as true or raise our degree of confidence in the hypothesis or rather assess the probability of accepting a false hypothesis if the test were run again and again.[1]

That inference to the best explanation and Bayesianism are compatible has been noticed before (see for instance Lipton, 2004: Chapter 7; Okasha, 2000). According to the position defended here, this is no accident. The two theories belong to different categories members of which play complementary roles. Inference to the best explanation is, despite its name, a theory of *evidence* that tells us what are the kinds of observations we should make, what are the kinds of tests we need to run in order to confirm or disconfirm a hypothesis. (As theory of evidence I will therefore refer to it as 'explanationism'.) It is silent about the types of inferences to be drawn from the evidence. Bayesianism is a theory of *induction* that tells us what inferences are warranted after the evidence has come to be believed. This theory is silent about what evidence is. A full account of learning from evidence requires both. Here I will look at theories of evidence first, then at theories of induction and finally at a number of hybrid theories.

Theories of Evidence

In this section I will look at two families of theories of evidence, instance theories and hypothetico-deductive theories. The former regards an instance of a hypothesis evidence for it, the latter, its entailments.

Instance Theories

According to the first family of theories of evidence, a state of affairs provides evidence for a general hypothesis if and only if it is an instance of the hypothesis. Here I will look at theories of evidence regarding two kinds of general hypotheses: simple subject-predicate hypotheses and causal hypotheses.

Simple natural laws

By 'simple natural laws' I refer to universally quantified statements that ascribe a property to a kind or substance such as 'All ravens are black' (Hempel, 1945) or 'All samples of the element bismuth melt at 271°C' (Norton, 2003). According to this first theory, evidence for such a generalisation is constituted by the instances of the generalisation.

Formulated in first-order logic, the account is subject to the famous 'ravens paradox'. The hypothesis 'All ravens are black' is logically equivalent to the hypothesis 'All non-black things are non-ravens'. Now, an instance of the latter hypothesis is a red shoe, and therefore observing a red shoe provides evidence for it. On the plausible assumption that if a state of affairs is evidence for one hypothesis, then it is evidence for a logically equivalent hypothesis, observing a red shoe provides evidence for the hypothesis 'All ravens are black', which is absurd (Hempel, 1945).

What is more important in the present context is that few interesting hypotheses in the social sciences have the form of a simple natural law. First, because of the high degree of biological, psychological and social variability, few claims are true of all tokens of a given type. Second, the theory restricts evidence to what is describable by 'phenomenal', that is, observable (such as 'black') or measurable (such as 'melts at 271°C'), predicates. To the extent that social science hypotheses have some generality, they tend to stem from explanatory theories, which involve theoretical predicates and require a different kind of evidence (see below).

Causal hypotheses

Despite these limitations, the idea that an instance provides evidence for its generalisation is behind a very influential set of principles for causal inference. There is a fundamental and critical distinction between sequences that are genuinely causal and those that are 'merely' accidental. To use a philosopher's stock example, smoking is said to *cause* cancer. On the other hand, many people who have nicotine-stained fingers will later develop cancer but having yellow fingers isn't a cause of cancer. The difference between the two is as important for explanation as it is for planning and policy. John's smoking may explain his cancer; stained fingers don't. And if John wants to avoid certain kinds of cancer, to stop smoking would be a good idea; avoiding yellow stains by wearing gloves wouldn't.

Building on the seminal work of Francis Bacon, John Stuart Mill developed five principles to distinguish between causal and accidental sequences: the methods of agreement, of difference, the joint method (of agreement and difference), of residues and of concomitant variation (Mill, 1843 [1874]: Book III, Chapter 3). Exemplarily, let us look at the first two in more detail.

The method of agreement seeks two sequences of situations in which the phenomenon of interest occurs such that they differ in every respect but one. Then, the factor that is followed by (or follows) is the cause (or the effect) of the phenomenon. Call ϕ the phenomenon of interest; f a factor and $\mathbf{X}$, $\mathbf{x}$, $\mathbf{Y}$, and $\mathbf{y}$ vectors of 'other circumstances' (where $X_i \neq Y_j$ and $x_i \neq y_j$ for all i, j and $X_i \in \mathbf{X}$, $Y_j \in \mathbf{Y}$, $x_i \in \mathbf{X}$, $y_j \in \mathbf{Y}$), then we can say schematically:

1 Method of agreement

$f\mathbf{X}$ is followed by $\phi\mathbf{x}$
$f\mathbf{Y}$ is followed by $\phi\mathbf{y}$
Therefore, f is a cause of ϕ.[2]

The method of difference, by contrast, seeks two sequences of situations, one in which the phenomenon of interest occurs and on in which it doesn't, that are exactly identical with respect to every factor save one, then that factor is the cause (or the effect) of the phenomenon. Schematically:

2 Method of difference

$f\mathbf{X}$ is followed by $\phi\mathbf{X}$
$\mathbf{X}$ is followed by $\mathbf{X}$
Therefore, f causes ϕ.[3]

The method of difference is the method of the controlled experiment. Both Mill's methods and the simpler instance theory constitute accounts of evidence because they tell us what the kinds of observations or tests are that one must take in order to support a hypothesis. It is true that both are associated with a rule of inference according to which the evidence warrants accepting the truth of the hypothesis – an inference rule I call 'categorical induction' (for Mill's case, see Mill, 1843). Norton (2003) therefore classifies the simple instance theory and Mill's methods under accounts of induction of the type 'inductive generalisation'.

However, the accounts are only accidentally wedded to this specific inference rule. There is nothing in the principles themselves that prevent using other rules such as probability updating or error correction (for a discussion of these rules of inference, see below).

Hypothetico-deductivism

The feature that unites the second family is that what makes a statement a statement

about evidence is its being entailed by a hypothesis and suitable auxiliary assumptions. The advantages of these theories over instance theories are immediate: any theory, using both predicates that refer to observables and those that refer to unobservables, can in principle be supported by evidence, not just generalisations. But there is also an immediate problem: few evidential statements are entailed by only one hypothesis. The usual case is that there are many mutually incompatible hypotheses, all of which entail the evidential statement. The main question for these theories of evidence is consequently how to discriminate between the different evidence-entailing hypotheses. I discuss two ways here, which I call eliminativism and explanationism, respectively.

Eliminativism

The most straightforward way to discriminate among evidence-entailing hypotheses is to devise tests or series of tests that eliminate all but one of the alternatives. This idea too goes back to Francis Bacon (see for instance Klein, 2003). The evidence relevant to a hypothesis is therefore constituted by the testable implications of the hypothesis at stake as well as those of its alternatives.

A large randomised controlled trial can serve as the paradigm because it eliminates many alternative hypotheses in one fell swoop. A hypothesis about the effectiveness of a new training programme, say, can be tested by dividing subjects at random into a treatment group (which receives the new training) and a control group (which receives the standard training) such that the distribution of other factors influencing performance is identical in the two groups. Then, if performance is on average higher in the treatment group than in the control group, it must be due to the new programme.

But it is not necessary that alternative hypotheses be eliminated by one and the same test such as a randomised trial. Michael Scriven proposes the following form of inference in the context of causal analysis in history (Scriven, 1966: 249–50; emphasis original):

> For in order to establish a causal claim on behalf of a factor what does the historian need? Merely evidence that his candidate was present, that it has on other occasions clearly demonstrated its capacity to produce an effect of the sort here under study (or there might be theoretical grounds for thinking it a possible cause rather than previous direct experience of its actual efficacy), and the *absence* of evidence (despite a thorough search) (a) that its *modus operandi* was inoperative here, and/or (b) that any of the other possible causes was present. If the event studied had a cause at all (which the historian assumes it did), then he may confidently assert that the residual condition is elected. This argument *proves* his claim – and it requires nothing the historian does not possess. The only general proposition that might be involved would be a list of the known possible causes of the kind of effect in question. Explanation proceeds by the elimination of possible causes …

The principal worry about elimination of alternative hypotheses in the social sciences is that is is frequently the case that methods such as the randomised trial are not applicable or applicable only in an attenuated form, and that there are too many possible alternatives not all of which can be ruled out. In medical research, a randomised trial is a powerful blinding device: neither patient nor doctor knows whether which of a number of treatments (including a placebo) is administered. Blinding in this sense is often not an option when a treatment is a training programme or some other social policy (Scriven, 2008).

Explanationism

If several theories or hypotheses entail the evidential statement, then that which best explains the evidence is confirmed according to this theory. Additional evidence is therefore provided by a fact about the theory: its relative degree of explanatoriness or explanatory quality.

One might argue that calling such a fact ‘evidence’ would be misleading. This is correct, but only on the first reading of the term ‘evidence’ as ‘data’ or ‘test result’. On the second reading of ‘evidence’ as ‘hint’, ‘sign’ or ‘indication’ (or even on the third reading

as 'proof'), there is nothing unusual about calling facts about a theoretical hypothesis evidence. At any rate, these facts are taken as evidence by proponents of explanationism.

A connection between evidence and explanation was already present in Hempel's (1965) account because according to his deductive-nomological theory of explanation, a hypothesis, if it is true, fulfils some additional criteria and, together with other statements entails the evidential statement, it *explains* the evidential statement. Turned around, one can say that according to this account, a statement is evidence for a hypothesis if and only if the hypothesis, if true, explains it. Explanationism adds a criterion to discriminate between competing potential theoretical accounts. In Gilbert Harman's words (Harman, 1965: 89):

> In making this inference one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis. In general, there will be several hypotheses which might explain the evidence, so one must be able to reject all such alternative hypotheses before one is warranted in making the inference. Thus one infers, from the premise that a given hypothesis would provide a "better" explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true.

This raises the question how to determine which of a given set of alternative hypotheses is 'best'. Many suggestions have been made: the simplest, the most unifying, the most detailed, that which confers most understanding on its user. This mode of reasoning is fairly common in theory-driven branches of the social sciences such as in parts of economics and sociology. In economics, for instance, a model is accepted as explanatory – or as more explanatory than an alternative – if it portrays a world that is *credible* (Sugden, 2000) or if it makes assumptions about structural features that can be found in a great range of economic phenomena; in other words, if the model is *unifying* (see Reiss, 2008: Chapter 6).

Explanationism is subject to an important objection sometimes called 'Hungerford's objection' (Lipton, 2004: Chapters 4 and 9) is that the various explanatory virtues such as credibility, simplicity, unifying power or mechanism are too subjective and varied to provide an acceptable ground for inductive reasoning.

An important role evidence plays in scientific and other controversies is that of an objective arbiter. If, say, one political party holds that minimum wages are an effective tool to provide a living wage for everyone and another that minimum wages are counterproductive because they destroy jobs, then such disagreements about apparently purely factual matters ('is *X* an effective strategy to promote *Y*?' when it is agreed that *Y* is a desirable state of affairs) should, in principle, be solvable on the basis of evidence. Inference to the best explanation reintroduces subjectivity through the back door because (a) there is no one generally accepted explanatory virtue; (b) there is no generally accepted schema that ranks or weighs the different explanatory virtues; (c) there are no objective criteria that determine whether and to what extent any given virtue applies to a given case – what's simpler (or more unifying or …) to one person isn't to another (on this latter point, see Norton 2003).

Contextualism

As can be seen, all theories of evidence that have been discussed so far have limited applicability. They work, to the extent that they do, only for a specific type of hypothesis and under favourable epistemic conditions. According to a position one may call evidential contextualism, this is to be expected (see for instance Reiss, 2008: Chapter 1; cf. Kincaid, 2007). Scientists' epistemic interests and their domains of investigation are too heterogeneous to subsume all kinds of clues that may indicate a hypothesis under one universal scheme or even a small finite set of schemes of limited generality.

According to evidential contextualism, it is context-specific background knowledge that informs scientists what kinds of techniques work in what domains and under what

conditions. At the general level no more can be said than that evidence is an observation or test result about which background knowledge entails that it is relevant to the assessment of a hypothesis of the given type and for the purpose at hand. There are numerous examples of the kinds of considerations that lead a (social) scientist to accept an observation or test result as evidence for different kinds of hypotheses and in the light of different kinds of purposes in the third section of this article.

Theories of Induction

Once we have evidence, what can we infer from it? This is the question addressed by theories of induction. In this section I will discuss three types of theory: categorical induction, two probabilistic theories and Norton's 'material theory of induction'.

Categorical Induction

What I call 'categorical induction' is the rule to infer the truth of the hypothesis from the evidence. Conjoined with the instance theory of evidence, we get enumerative induction or the 'more of the same' rule of inference. Conjoined with the first form of hypothetico-deductivism, we get eliminative induction and conjoined with its second form, inference to the best explanation or abduction.

Enumerative induction

This rule prescribes to infer from a finite set of observed instances to the corresponding generalisation. For example,

> Raven 1 is black
> Raven 2 is black
> …
> Raven *n* is black
> Therefore, all ravens are black.

One problem is that the account is vague: it does not say how many instances must be observed in order to warrant inferring the truth of the generalisation. Another problem is that the rule itself is underspecified unless supplemented with an account of natural kinds that provides limits for the types of predicate that are admissible for being quantified over. Nelson Goodman's famous 'grue paradox' illustrates the problem.

Goodman's paradox, in short, is as follows. The evidential statement 'All emeralds that have been observed so far have been green' confirms the generalisation 'All emeralds are green'. Now, we can define a new predicate, 'grue' as 'green if examined before t or blue otherwise'. Thus, before time t, we have exactly parallel evidence that all emeralds are green and that all emeralds are grue. But obviously only one of the two hypotheses can be true. The question, then, is which one we are warranted to infer. Some (e.g. Quine, 1969) have argued that 'green emerald' represents a natural kind while 'grue emerald' doesn't – and therefore that we are warranted to infer the former hypothesis but not the latter.

And it is not necessary to introduce artificial predicates such as 'grue' to make that point. A more scientific example due to John Norton can be used similarly. Consider the following inferences (Norton, 2003: 649):

> Some samples of the element bismuth melt at 271°C.
> Some samples of wax melt at 91°C.
> Therefore, all samples of the element bismuth melt at 271°C.
> Therefore, all samples of wax melt at 91°C.

The obvious difference between the two arguments is that 'bismuth' refers to a chemical element – a type of natural kind – whereas 'wax' names a variety of different mixtures of substances. Using simple enumerative induction as an inference rule is successful in the former case because of a known fact about chemical elements: elements are generally uniform in their physical properties. No such fact is true of the different mixtures of hydrocarbons that are jointly referred to as 'wax'.

Of course, restricting the inference rule to hypotheses regarding natural kinds is epistemically not helpful. Natural kinds are

just those kinds certain properties of which are uniform across all instances, Knowing that already presupposes that the inferential problem has been solved.

Eliminative induction

Alexander Bird calls this rule inference to the only explanation (in contrast to inference to the best explanation, see below) and describes it thus (Bird, 2007: 242):

> By Inference to the Only Explanation (IEO) I intend something quite specific, that at the end of inquiry we can be in the position to infer the truth of some hypothesis since it is the only possible hypothesis left unrefuted by the evidence. It is the form of inference advocated by Sherlock Holmes in his famous dictum 'Eliminate the impossible, and whatever remains, however improbable, must be the truth.'

There are three main worries that beset this form of induction. First and foremost is the idea that theoretical hypotheses are always underdetermined by the evidence available at a certain point in time. This is the Duhem–Quine underdetermination thesis. There are good reasons to believe that the thesis in its most general form – any theoretical hypothesis is always underdetermined by all available evidence – is false (see for instance Norton, 2008). In many actual scientific cases, all *plausible* hypotheses but one could be eliminated. The just mentioned article by Bird reconstructs Semmelweis' discovery of the cause of puerperal fever along these lines. But in the social sciences we often face situations in which the available evidence *de facto* underdetermines the choice of theoretical hypotheses. The list of possible and even plausible alternatives can be very long indeed, potentially open ended. Thus, the 'problem of confounders' (Steel, 2004) is a real obstacle to social-science research.

There are two further problems but they are less specific to social science. One is the question whether the list of plausible alternatives contains the true hypothesis. Eliminative categorical induction is only guaranteed to result in the true theory if it does. But, as Bas van Fraassen reminds us, the rule 'infer the truth of the hypothesis that is the only one to be consistent with the evidence' (1985: 143), '… is a rule that selects the best [supported] among the historically given hypotheses. We can watch no contest of the theories we have so painfully struggled to formulate, with those no one has proposed. So our selection may well be the best of a bad lot'. The final criticism is whether the evidence requires an explanatory hypothesis at all. Perhaps some facts are just brute.

Inference to the best explanation

The proponent of inference to the best explanation or abductive reasoning infers the truth of a hypothesis from two considerations: (a) the hypothesis *explains* the evidence; (b) among the evidence-explaining alternative hypotheses it is the one that scores best on some scale of explanatory merit. For Peter Lipton, for example, the 'loveliness' of an explanation is the relevant criterion. Hence, the loveliest explanatory hypothesis is inferred to be true, according to this rule.

This inductive schema is subject to what has been called 'Voltaire's objection' (Lipton, 2004: Chapterss 4 and 9). The objection denies the connection between goodness and truth. It asks, why should the world be simple or intelligible or lovely? In the context of the social sciences the link between explanatory 'loveliness' and truth often seems particularly tenuous. In economics, for example, a 'lovely' explanatory model portrays agents as perfectly rational, uses equilibrium concepts to solve equations and is mathematical in nature. There is little plausibility in the idea that considerations such as these should be a reliable guide to truth (see Reiss, 2008: Chapter 6).

Probabilistic Theories

Probabilistic theories ascribe a probability to the hypothesis and understand evidential support in terms of probabilistic relations.

Bayesianism

Standard Bayesianism combines an interpretation of probability as subjective degrees of

belief with Bayes' rule – itself a theorem of probability theory – and an interpretation of belief-updating as confirmation to yield a schema for making inductive inferences.

Aside: Five interpretations of probability

The five major interpretations of probability are: classical, logical, subjective, frequency and propensity. The *classical* theory holds that probability is the ratio of the favourable cases to the total number of equally possible cases. For instance, the probability of the event 'rolling a number larger than three with a fair die' is (1 + 1 + 1)/6 = 1/2. According to the *logical* interpretation, too, all possible states of affairs are assigned probabilities but it relaxes the requirement of equal weights. A probability measure assigns numbers to so-called state descriptions, which describe all individuals in a universe in maximum detail.[4] According to *subjectivists* such as Bayesians, probabilities are constraints a rational agent lays upon the degrees to which he holds a belief. Probability expresses the confidence with which an agent holds a belief. *Frequentism* identifies probability with the frequency of favourable outcomes in a (finite or infinite) reference class (e.g. the frequency of 'heads' in a finite series of tosses of a coin or a hypothetical infinite series). Finally, according to *propensity* theorists, probability is a physical disposition of a chance set up to generate outcomes (such as the tendency of an atom to decay within a certain amount of time).

Bayes' theorem has numerous forms (see Howson and Urbach, 1993: Chapter 2); for an arbitrary hypothesis h, its negation $\sim h$ and evidential statement e, it can be expressed as follows:

$$P(h \mid e) = P(h)/[P(h) + (P(e \mid \sim h)/P(e \mid h))P(\sim h)] \quad (29.1)$$

where the expression $P(e \mid \sim h)/P(e \mid h)$ is called the 'likelihood ratio'. Since $P(\sim h) = 1 - P(h)$, $P(h \mid e)$, the posterior probability of the hypothesis given the evidence, is a function of its prior probability $P(h)$ and the likelihood ratio.

Together with the interpretation of probability as subjective degree of confidence or belief in a hypothesis and the idea that updating entails confirmation, Bayes' theorem entails an inference rule: upon coming to believe the evidential statement, update your degree of belief in the hypothesis in accordance with (29.1). Evidence in favour of (against) a hypothesis increases (decreases) the degree of belief in the hypothesis.

To illustrate how Bayes' theorem works as an inference rule, consider a medical example. Suppose that a diagnostic test is 99 per cent accurate; that is, it gives a correct test result in 99 per cent of cases, both positive (when the disease is present) and negative (when the disease is absent). If the test gives a positive result, what is the probability that the person actually has the disease? Let h = 'person has disease' and e = 'test result is positive'. The quantity we would like to infer is the posterior probability $P(h \mid e)$, the probability that the person has the disease *given* that the test result is positive. The likelihood ratio is $P(e \mid \sim h)/P(e \mid h) = 1\%/99\%$. Let the h's prior probability, in cases such as this called the base-rate, be 1/10,000 (i.e. one person in ten thousand has the disease). Then, by (29.1):

$$P(h \mid h) = .0001/[.0001 + (.01/.99)*.999] \approx .00981$$

Because of the low base-rate, the probability that the person taking the test has the disease, despite the positive test result, is below 1 per cent. Nevertheless, a positive test result provides evidence for the hypothesis because $P(h \mid e) \approx 1\% > .01\% = P(h)$.

Several criticisms have been levied against the Bayesian inductive rule, two of which I want to discuss here. According to the Bayesian, a hypothesis' probability assessment after coming to believe the evidence depends, as we have seen, on two factors: the hypothesis' prior probability and the likelihood ratio. The first criticism finds fault with

prior probabilities; the second is suspicious that evidence enters the inference *only* via the likelihood.

In ordinary probability theory, probabilities are defined over events or outcomes or a sample space. Making probability statements entails that things could be different than the way they actually are. For instance, saying that the probability of rain today is 90 per cent means that it might either rain or not (and that it should rain on 90 per cent of days like today). Scientific hypotheses, by contrast, are either true or false (or, if one prefers, empirically adequate or not or reliable or not), they are not 'probable' in the sense that rain is. If there were probabilities, this would imply that there are many worlds, in which case the probability of a hypothesis could measure the frequency of worlds in which the hypothesis is true. But there is only one world – and thus, there are no priors (Mayo, 1996: Chapter 3).

The other criticism can be illustrated by the debate regarding stopping rules. A stopping rule specifies when to stop collecting new data. To test whether a coin is fair, for instance, one might toss it 20 times, record the number of heads and tails (say, 8 and 12) and then assess whether the specific outcome was more likely if the coin was fair or if it was unfair. Alternatively, one might continue tossing the coin until one has recorded eight heads or a ratio of two heads to three tails. Intuitively, it should matter a great deal how the test is set up. For instance, it should matter whether a specific outcome was likely or not given how the procedure was designed. Bayesianism entails that to the assessment of the fairness of the coin these considerations should not play a role. Another way of putting it is that Bayesianism is only sensitive to the actual outcome (the actual series of 8 heads and 12 tails) not also to the outcomes the test could have produced but did not.

Likelihoodism

The likelihood view is essentially Bayesianism without the priors. An attractive feature of it is that it regards evidence always as relevant to a hypothesis relative to an alternative. The law of likelihood states that evidence e supports hypothesis h_1 over hypothesis h_2 if and only if $P(e \mid h_1) > P(e \mid h_2)$ (see Hacking, 1965).

As in Bayesianism, a positive test result is evidence for the hypothesis that a person has the disease, even if the posterior probability might be low. Using the same numbers as above and defining h_1 as the hypothesis 'John has the disease' and h_2 as its negation, the evidence 'positive test result' supports h_1 as:

$$P(e \mid h_1) = .99 >> .01 = P(e \mid h_2).$$

Likelihoodism is an account of evidence that addresses the question, 'Which of two hypotheses is better supported by the evidence?' That this can lead to counterintuitive results is illustrated with example due to Ian Hacking (Hacking, 1972: 136): 'We capture enemy tanks at random and note the serial numbers on their engines. We know the serial numbers start at 0001. We capture a tank number 2176. How many tanks did the enemy make? On the likelihood analysis, the best supported guess is: 2176.' That is, after capturing that one tank with the number 2176, the hypothesis that the number of tanks the enemy made is just that number is better supported than any other hypothesis. As likelihoods are the only way evidence enters reasoning, this account is subject to the criticism regarding stopping rules in the same way as Bayesianism is.

Naturalism

Naturalism asserts that the best place to look for insights regarding inductive rules is science itself. It is suspicious of substantive philosophical claims of great generality that are made independent of the details of specific scientific practises. Here I will look at two examples: Norton's 'material' theory of induction and the error correction perspective.

Norton's 'Material' Theory of Induction

All above mentioned families of theories of induction purport to have universal range,

that is, they are thought to apply to every domain of inquiry, independently of the more specific facts true within the different domains. In a recent paper John Norton has challenged the feasibility of the general project behind these 'formal' theories of induction as he calls them (Norton, 2003). Unlike deductive schemata of inference such as modus ponens or universal instantiation, which do enjoy universal validity, there is no inductive schema that has not been subject to criticism and counterexample. All examples of inductive rules discussed above are cases in point.

Hence, according to Norton, what licenses the inference is not the *form* of the inductive schema – as that is the same in instances where it works and where it doesn't – but rather particular material facts true of the situation in which the inference is made. Norton shows that all formal inductive schemata work where they do because of such material facts. Thus, inductive inferences derive (Norton, 2003: 648):

> their license from facts. These facts are the material of the inductions; hence it is a 'material theory of induction'. Particular facts in each domain license the inductive inferences admissible in that domain – hence the slogan: 'All induction is local.' My purpose is not to advocate any particular system of inductive inference. Indeed I will suggest that the competition between the well established system is futile. Each can be used along with their attendant maxims on the best use of evidence, as long as we restrict their use to domains in which they are licensed by prevailing facts.

Material facts, according to this theory, license not only that inductive inferences are made from evidence but also what specific types of inferences can be made.

Error correction

Inferences from evidence to a hypothesis are subject to a variety of errors able to invalidate conclusions. As an example that is highly relevant in the context of the social sciences, consider a correlation between two variables X and Y as evidence for a causal hypothesis. That correlation does not entail causation is well known. If a correlation is nevertheless taken as evidence for a causal claim – and it certainly provides a clue that two variables may be causally related – other sources of correlation must be controlled for if the hypothesis is to be inferred reliably. If the hypothesis of interest is 'X causes Y', one first wants to rule out reverse causation from Y to X as well as common factors that influence both variables. In addition there are numerous non-causal sources of correlation: sampling error, measurement error, non-stationarity and other statistical properties of the variables, mixing, variables that are conditioned on common effects and so on. A hypothesis can be inferred reliably to the extent that these sources of error have been controlled successfully (Reiss, 2008: Chapter 1; cf. Schickore, 2005; Hon, 1989, 1995).

There is no general account of error that is independent of the type of hypothesis, specific domains of science and the purposes to which the hypothesis is put. This is the main difference between this account and eliminative induction. There is no requirement that either the hypothesis of interest or a potential inferential error explain the evidence. A hypothesis about the future value of a variable, for instance, does not explain whatever evidence one might have in its favour, nor does a descriptive hypothesis about, say, the inflation rate, explain why an index number has such and such a value (see sections on Index Numbers and Expert Political Judgement, respectively). The source of information about potential errors in inference lies, rather, in context-specific background knowledge about the domain of investigation.

There are two main problems with naturalism. First, it is hardly a *theory* of induction. Theories in the philosophy of science do not only aim to show what is the common logic behind a scientific practise such as scientific explanation, inference, measurement or experimentation but also to *explain the rationale* behind these practises. Norton's account does neither. He denies that there is a universal logic of induction. The second

naturalistic account appeals to vague 'logic of controlling for known errors' but neither account provides a justifying rationale for why a given inferential practise can be expected to work or why it is rational to draw inferences in the way described. These accounts therefore lack normative power. If one asks, say, why a certain methodology such as that of the randomised trial is as successful as it is and one hears that this is due to specific facts about the 'physical probabilities of the randomizer' (Norton, 2003: 655), one hasn't been answered.

Second, the different formal theories differ dramatically in both their informational requirements or inputs as well as their outputs. Bayesianism, say, requires a prior probability and likelihoods and yields a posterior probability. By contrast, the error-*statistical* approach denies that prior probabilities attach to hypotheses. Naturalism has no resources to inform us about where and when either inference rule can and should be used. This is important as the different rules yield different results *even in cases where the material facts are undisputed* (Steel, 2005a). An appeal to scientific practise isn't informative when scientists themselves are divided about what inferences are licensed by a situation.

Hybrid Theories

Hybrid theories provide resources that not only allow us to classify a piece of information such as an observation or test result as evidence, they also contain rules of inference. These accounts do not simple conjoin a theory of evidence with an inference rule. Rather, the two aspects are parts of an integrated whole. In this section I discuss two such theories, the error-statistical account and Achinstein's theory.

Error Statistics

The error statistical account develops classical statistical testing of Neyman and Pearson into a full-fledged philosophical theory of evidence and induction. As a theory of evidence, it regards an observation evidence for a hypothesis to the extent that it has been produced by a test procedure that would have made it very unlikely that the observation would have been produced if the hypothesis were false. More precisely (Mayo, 2000: S198; emphasis original): 'Data *e* produced by procedure *T* provides good evidence for hypothesis *H* to the extent that test *T severely passes H* with *e*'.[5] Hypothesis *H* passes a *severe test* with *e* if (i) *e* fits *H* (for a suitable notion of fit or 'distance') and (ii) the test procedure *T* has a very low probability of producing a result that fits *H* as well as (or better than) *e* does, if *H* were false or incorrect.

Thus, the approach requires that hypotheses be subject to a test that is as stringent as a randomised controlled trial (for a detailed discussion, see section on Evidence-Based Policy). A randomised trial eliminates all sources of error in one fell swoop. Hence, it will 'pass' the hypothesis if it is false only for statistical reasons, because of sampling error. This error can, however, be controlled by the procedure's 'error-probabilities'.

There are two error probabilities. Type-I errors consist in rejecting the null hypothesis (usually the hypothesis that there is no treatment effect) when it is in fact true. It is controlled by choosing the significance level of the test. Type-II errors consist in not rejecting the null when it is in fact false. It is controlled by designing the test such that it has high power, which is one minus the probability of a type-II error. Significance level and power aren't entirely independent, however. If sample and effect size are given, fixing the significance level determines power and vice versa. Thus, at a chosen level of significance (and assuming that the effect size cannot be manipulated), the power of the test can only be increased by increasing the sample size.

Error statisticians Deborah Mayo and Aris Spanos add a third error probability to these, that of 'attained power' or 'severity' (Mayo and Spanos, 2006). It is a post-data measure

and therefore sensitive to the sample realisation (whereas significance and power are pre-data and independent of the realisation). To measure severity, one must define a value for discrepancy from the null one deems as substantially significant (for instance, a treatment effect size) Severity then defined as one minus the probability that a test result like the one obtained occurs if the discrepancy is in fact larger.

What we infer from a test about the hypothesis therefore depends on what we deem as scientifically significant. Suppose that a new training programme is under scrutiny. The null hypothesis says that it is ineffective (relative to the current best programme, say). Testing the proposition by statistical means does not by itself allow us to draw an inference. If the test yields an insignificant result (let us say), we cannot simply infer that there is no treatment effect. In addition we have to specify the type of inference we wish to draw in terms of a distance from the null. Then, once the data is in, we can calculate the probability of the test producing data like these if the distance is in fact greater, that is, if there is in fact a (greater) treatment effect.

Intuitively, this makes a lot of sense. If a certain result is produced that is statistically insignificant, it is one thing to conclude 'there is (probably) no effect' and quite another 'there is (probably) no *large* effect'. Accordingly, the severity with which the test passes these two hypotheses differs. The same is of course true when the result is significant. Then, given the result, the larger the discrepancy (i.e. the treatment effect) one wishes to infer from this result, the lower the severity.

Finally, the inference we draw is about the frequency of achieving a certain test result were the test repeated many times, just as in classical statistics. Thus, we do not infer a hypothesis (as in categorical induction) or a degree of belief in a hypothesis (as in Bayesianism) but rather a claim about a probability of a certain test result. That is, the probability claims of the error-statistical approach attach to test procedures, not to the hypothesis of interest. The confirmation of hypotheses is therefore of a purely qualitative nature.

Construed as a theory of evidence, the error-statistical approach is exceedingly narrow as it can only be used to test statistical hypotheses. To be sure, statistical inference is an important part of research in all areas awash in data. But it is just one part. To infer a causal claim from an experiment we do want to know whether the observed correlation is real or due to chance. But there are many other sources of error that we want to control – measurement error, confounding, non-causal sources of correlations and what have you. The error-statistical approach has no answer to these.[6]

Achinstein's Theory

Peter Achinstein combines a form of explanationism as theory of evidence with a form of Bayesianism as inductive rule. He first distinguishes various concepts of evidence: subjective, ES-, potential and veridical (Achinstein, 2001: Chapter 2). Essentially, an agent has *subjective* evidence *e* for *h* if she believes *e* and takes it as a reason to believe *h*. ES refers to an agent's 'epistemic situation'. In a specific epistemic situation *C*, *e* is *ES-evidence* that *h* if, in *C*, *e* is a good reason to believe *h*. *e* is potential evidence in *h* if it is a good reason to believe *h simpliciter* and *e* is veridical evidence that *h* if *e* is potential evidence that *h* and *h* is true.

Achinstein thinks that veridical and potential evidence are the concepts most relevant for scientific practise, and most of his work is dedicated to explicating those concepts. He defines (Achinstein, 2001: 170):

> PE. *e* is potential evidence that *h*, given *b*, [if and[7]] only if
>
> 1 P(there is an explanatory connection between *h* and *e* | *e*&*b*) > 1/2
> 2 *e* and *b* are true
> 3 *e* does not entail *h*

where *b* signifies background knowledge. By 'explanatory connection' Achinstein means

that it is either the cases that the hypothesis explains the evidence, that the evidence explains the hypothesis or that a common factor explains both. This condition is supposed to rule out classical counterexamples to 'high probability' accounts of evidence and confirmation. Achinstein's example (p. 149):

> *h*: Michael Jordan will not become pregnant.
> *e*: Michael Jordan eats Wheaties.
> *b*: Michael Jordan is a male basketball star.

On a simple high probability account, *e* is evidence that *h* because the probability that *h* given *e* is high. Achinstein assumes that in this case *e* and *h* are not explanatorily connected: neither does his eating Wheaties explain that he won't become pregnant nor vice versa, nor is there a common factor that explains both.[8] Eating Wheaties is explanatorily irrelevant for not becoming pregnant.

Achinstein uses the general requirement that *h* and *e* must be explanatorily connected rather than that, say, *h* explain *e*, because causes can provide evidence for effects, effects for causes and joint effects of common causes for each other. Thus, taking a potent medicine can be evidence for relief, relief for taking medicine and the drop in the barometer reading can be evidence for the storm.

To require that there be an explanatory connection between evidence and hypothesis is plausible but too strong. If evidence is to be a mark or a symptom of the truth of a hypothesis it is enough that there be a correlation between the states of affairs hypothesis and evidential statement express, there need not be causal or explanatory relation. A widely discussed case involving a 'spurious correlation' demonstrates this. Elliott Sober describes his case thus (Sober, 1987[1994]: 161–162 [quoted from Sober, 2001: 332]):

> Consider the fact that the sea level in Venice and the cost of bread in Britain have both been on the rise in the past two centuries. Both, let us suppose, have monotonically increased. Imagine that we put this data in the form of a chronological list; for each date, we list the Venetian sea level and the going price of British bread. Because both quantities have increased steadily with time, it is true that higher than average sea levels tend to be associated with higher than average bread prices. The two quantities are very strongly positively correlated.

Now, if there is a strong correlation between the two quantities, we can use, say, the fact that one quantity is very high as evidence for the hypothesis that the other is high as well. But there isn't any explanatory connection between the two. High bread prices do not explain high sea levels, nor do high sea levels explain high bread prices and nor is there a common factor that explains both. Time can be used to predict whether the quantity is high or low but it does not explain why this is so. Or, to take a case due to Jossi Berkovitz (as described by Dan Steel, 2005b: 19): '… imagine two slot machines constructed entirely independently of one another but which, coincidentally enough, have precisely the same initial conditions and internal mechanics'. Here we are facing a brute correlation (or a set of such correlations) that once more has no explanation. And again, we can use the state of one machine to predict – provide evidence for – the state of the other.

That correlations rather than explanations are required as evidentiary relationship is also shown by the reverse case, where two quantities have an explanatory connection without a correlation being induced. A philosophers' stock example regarding the connection between correlation and causation concerns the relation between birth control pills and thrombosis (originally due to Hesslow, 1976). Birth control pills cause thrombosis via one route but they also prevent thrombosis by preventing pregnancies, as pregnancies are themselves a cause of thrombosis. Depending on the actual frequencies, the probability raising and lowering channels might just cancel each other out so that the probability of developing thrombosis is the same whether or not a given woman takes the pill. Knowing that, we should not take facts about oral contraceptives as evidence for hypotheses about the likelihood of contracting thrombosis. Nevertheless, there is an explanatory connection. For instance,

the reason for a given occurrence of thrombosis might lie in the woman having taken the pill.

The seemingly arbitrary requirement that the probability of there being an explanatory connection be greater than .5 stems from Achinstein's absolutist concept of evidence. A good reason to believe *h* cannot also be a good reason to believe not-*h*. Hence the probability given the evidence and background knowledge must at least be .5. Further, Achinstein includes the third condition because does not want evidence and hypothesis to be too 'close' to each other. The drop in the barometer reading is not evidence for the change of the barometer reading.

As inferential rule, Achinstein uses Bayesian updating, which makes him a Bayesian of sorts. The main difference between his account and standard Bayesian is his 'objective epistemic' interpretation of probability (Achinstein, 2001: Chapter 5). Standard Bayesianism interprets probability as degree of belief. Apart from adhering to the axioms of probability theory, there are no constraints on what a subject ought to believe.[9] Achinstein defines probability as 'degree of reasonableness of belief'. It is therefore not a measure of how strongly a person believes in a proposition but rather one of the quality of the reasons for holding a belief. Further, it is not subjective in the sense of being relative to a particular agent. The degree of reasonableness of a certain drug producing relief may be .8 even though no single agent holds this belief.

Apart from this difference in interpreting probability, Achinstein is a Bayesian. In particular, he must assume hypotheses (and evidence) to have prior probabilities, which makes his account vulnerable to the same objection regarding priors as standard Bayesianism.

SOURCES OF EVIDENCE

As indicated in the Introduction, there is an enormous variety of evidence-generating methods or 'sources' of evidence, both across all sciences as well as within a single science. It is impossible to review all the methods used in the social sciences here, or even all those one can find in a single social science. In what follows, I will therefore present a highly selective partial overview. The selection is guided partly by considerations of scientific importance and philosophical interest but partly also by my expertise. I group the methods into three categories: sources of evidence for (a) descriptive claims; (b) explanatory claims; and (c) policy claims.

Descriptive Inference

This type of inference and the associated methods are frequently ignored by philosophers but they are all the more important in social science. Hardly any property that is of interest from a social science point of view is immediately observable. In order to establish facts, even purely descriptive facts about a society, the investigator has to make inferences on the basis of new immediate observations, already established facts and background knowledge. I focus on two examples here: participant observation (used, for instance, in anthropology, communication studies, criminology, social psychology and sociology) and index numbers (used mostly in economics). In a sense, participant observation and index numbers represent the two extremes of the same spectrum. The participant observer is an actor who, by immersing herself in the culture she researches, becomes an expert in that culture and as such makes informed *judgements* about whether this or that fact obtains. When establishing facts by means of index numbers, the aim is to reduce expert judgement to a minimum by standardising procedures. To be sure, measuring quantities such as inflation or unemployment too requires judgements at various points, even when procedures are standardised. For example, a government statistician must judge whether the goods he finds in a chosen supermarket are indeed

comparable to the goods chosen previously, and if not, what type of adjustment procedure to use. Similarly, Bureau of Labor surveyors must interpret the responses of households when measuring unemployment.

Participant Observation

A key idea of participant observation is that the researcher occupy a role within the group she observes and its aim, at least when done in cultural anthropology, is to produce an ethnography. Typically, participant observation involves (DeWalt and DeWalt, 2002: 4):

> living in the context for an extended period of time;
> learning and using local language and dialect;
>
> - actively participating in a wide range of daily, routine, and extraordinary activities with people who are full participants in that context;
> - using everyday conversation as an interview technique;
> - informally observing during leisure activities (hanging out);
> - recording observations in field notes (usually organized chronologically); and
> - using both tacit and explicit information in analysis and writing.

There are two main forms of the technique: overt and covert participant observation. In the former case, the observed group both knows and permits the researcher to participate and conduct her investigation, which has advantages and disadvantages. Advantages include easier access to the group as a whole and subgroups (if one is allowed to participate!) as well as easier recording of the observations made; the main disadvantage is that as a result of knowing to be observed the group may change its behavioural patterns (the so-called 'observer effect'). Covert participant observation is carried out secretly, without the group's knowledge and permission. Apart from obvious ethical worries, disadvantages include the danger of losing objectivity on the researcher's part and greater difficulties in recording data. The advantages are that access to groups that would not normally allow it can be gained and that there are greater chances of avoiding the observer effect.

A similar trade-off obtains at the level of the degree of involvement of the individual participant observer. One end of the spectrum is occupied by the researcher who 'goes native' never to return from the field. His immersion is complete but just as complete is his loss of objectivity and detachment, and he obviously relinquishes the aim of producing a scientific outcome. At the other extreme are researchers who aim to keep active involvement at a minimum but thereby also forfeit the goal of gaining entry to inside knowledge. One important source of ethical conflict in this area is the question to what extent a researcher should intervene if she finds an observed practise objectionable.

Another problem is that involved research can be intrusive, which raises concerns about the privacy of the observed groups; but there is often no other way to access information of this kind.

A further trade-off besets the description of the observed social practises. Traditionally, social 'facts' used to be reported as if valid for all (places and) times: 'Social group G engages in practise ψ'. After the 'reflective turn', however, reports resembled more what in philosophy is called a protocol sentence: a statement about the observation of a concrete event, along with details about when, where and how the observation was made. This mode of recording facts has been criticised as subjective and even as overly indulgent in personal details of the researcher. On the other hand, it reduces the risk of hasty generalisation and unwarranted inference.

Index Numbers

Index numbers are widely used in economics in order to estimate quantities of interest such as the price level, inequality or wellbeing. Suppose for example we would like to assess whether the price level in the following two-by-two toy economy has increased, decreased or stayed put as illustrated in Table 29.1.

Only if all prices change at the same rate, an unequivocal answer can be given. If not, as in the example, some weighted average must be drawn, and different methods of averaging or

Table 29.1 A two-by-two toy economy

	Price cocoa	*Quantity cocoa*	*Price cloves*	*Quantity cloves*
Year 1	100	3	50	2
Year 2	50	2	100	3

Source: Reiss, 2008: 68

aggregating the raw data give different, sometimes widely disparate, results. Table 29.2 shows five different indices, computed for the above data.

Index numbers are a particularly clear case showing that philosophical theories of evidence have few resources to help with concrete scientific problems concerning evidence. Suppose our index number, say, a Laspeyres index, computes the inflation rate in a region and period to be 3 per cent. Call this our evidence *e*. Then suppose our hypothesis is that the inflation rate in that region and period is indeed 3 per cent. There is no explanatory connection between the evidence and the hypothesis. The evidence is a mathematical construct, and there is no good sense in which it is caused by (even less so, causes) the quantity of interest. Nor is there a good sense in which observing the evidence should lead us to revise our belief in the hypothesis – unless further assumptions are made (and then it is these further assumptions that justify the inference). There is also no good sense in which the 'test' (computing the index number) passes the hypothesis 'severely'. What is the probability that the index yields this result if 'true' inflation were different from 3 per cent?

Rather, whether or not the datum '3 per cent' is evidence for our hypothesis depends on considerations of the following kind. A Laspeyres index – a *price* index – gives exact results only if nothing but prices change in the economy. If other things change, for instance traded quantities, the answer is ambiguous unless further assumptions are made. If we assume that consumers respond to price changes by adjusting expenditures (relatively more expensive goods are substituted by relatively cheaper goods), it can be shown that the Laspeyres index *over*states inflation. To get rid of this 'substitution effect' one can compute a so-called 'superlative index', of which the geometric mean between Laspeyres and Paasche index is an example.

Table 29.2 Different indices

	Arithmetic	*Geometric*	*Harmonic*	*Laspeyres*	*Paasche*
Rate of change	25%	0%	–20%	–13%	14%

Source: Reiss, 2008: 68

Price and quantity changes are not, however, the only changes between two periods. Tastes, environments, the quality of the exchanged goods as well as the range of available goods may change too. In each of these cases decisions must be made about how to adequately incorporate a source of change into the index. The index-number purpose will guide these decisions. For instance, to measure consumers' cost-of-living, it makes sense to include mortgage payments in the budget. By contrast, if the purpose is to test monetary theory, mortgage payments should be excluded as they are directly proportional to interest rates, which play an important explanatory role in that theory.

Explanatory Inference

Accurate descriptive inference is an important goal of social science in its own right. It also plays a preparatory role for further inferences regarding the explanation of social phenomena. There are numerous models of explanation in the social sciences but the causal model is currently the dominant, and I will focus on causal inference here.

Qualitative Comparative Analysis

Above, we have already looked at Mill's methods of causal inference. Squarely in the Humean tradition, Mill understood causation to be a kind of complex regularity. To him, a cause was a insufficient but non-redundant

part of an unnecessary but sufficient, short INUS, condition (this analysis is due to John Mackie, see Mackie, 1974). To define a cause as INUS condition says three things.

> Any cause is followed by its effect only if certain enabling conditions are present and disturbing factors absent.

1 For any effect, there are many alternative sets of causes that may precede it.
2 The relationship between cause and effect is invariant; that is, when the right causal conditions are in place, the effect must follow and vice versa.

Qualitative comparative analysis (QCA), developed by the sociologist Charles Ragin (see for instance Ragin, 1998) builds on the understanding of cause as INUS condition and makes use of it in drawing causal conclusions from comparing a small number of cases. It has been applied to to fields as wide-ranging as sociology, political science, economics and criminology (for a full list of applications, see the bibliographical database at www.compasss.org). QCA aims to overcome the problem of small sample sizes by making the maximum possible number of comparisons among the sampled units.

The method identifies causes of phenomena of interest (e.g. ethnic political mobilisation among Western European minorities) by first arranging the all observed instances (in this case, minorities) in a table and determining whether or not the phenomenon is present. Then a list of factors (in this case, size, linguistic ability, wealth relative to core region and population growth) is constructed and it is noted whether each factor is present or absent. A factor is judged to be a cause whenever it is a member of a group such that that group of factors is always associated with the phenomenon of interest and no subgroup is always associated with the phenomenon – in other words, if it is an INUS condition.

The analysis of causation in terms of INUS conditions is deficient, as Mackie himself understood. In his famous example it is the sounding of the Manchester hooters at 5.00 p.m. that is shown to be an INUS condition for the Londoners to leave work shortly thereafter, but of course the Londoners do not leave the factory *because* of the sound of the Manchester hooters (see Mackie, 1974: 81–4). Nevertheless, Ragin's account demonstrates how regularities of a certain kind can constitute – defeasible – evidence for causal hypotheses. The most likely source of deficiency is the omission of common causes. As Mackie's hooters example shows, omitting a common cause – in his case, it being 5.00 p.m. – leads one to misinterpret what is in fact joint effect as a cause. When causes operate indeterministically – in the social sciences a possibility one should not exclude *a priori* – the full set of causal conditions does not have to be sufficient for its effect, which also makes the application difficult in this area.

Causal Modelling

Whereas QCA is based on (or can be explicated with) the Mill-Mackie analysis of 'cause' as an INUS condition, the various approaches to causal modelling relate to the probabilistic theory of causation, according to which causation is a specific form of correlation or probabilistic dependence.[10] The most popular form of causal modelling is currently that of the so-called Bayesian networks or short Bayes' nets.

A Bayes' net is a directed acyclic graph or DAG with an associated probability distribution. A graph is a set of vertices (representing *variables*) and a set of edges connecting the vertices (representing *relations* among the variables). A graph is directed when all its edges are directed and acyclic when there are no directed cycles. There is assumed to be a joint probability distribution over the variables. If the graph is Markov, it can be used to represent certain kinds of probabilistic independencies among the variables. For instance, in both DAG1 X1 and X3 are independent conditional on X2 and in DAG2 X2 and X3 are independent conditional on X1 as shown in Figure 29.1.

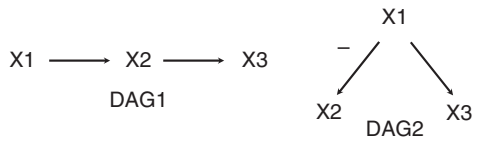

Figure 29.1 Two directed acyclic graphs (DAGs)

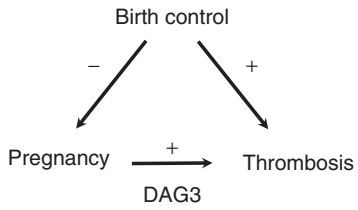

Figure 29.2 A potential counterexample to FC

By themselves, DAGs are abstract mathematical objects that could be used to merely store probabilistic information efficiently. From the point of view of causal inference, they become interesting when vertices are interpreted as causal factors and edges as causal relations. DAG1, for instance, can be used to represent the causal chain 'obesity causes diabetes causes heart attacks', and DAG2 the common-cause structure 'drop in atmospheric pressure causes change in barometer reading and storm'.

A causal Bayes' net is assumed to satisfy the causal Markov condition (CMC), according to whichCMC is a variable X is independent of all other variables in a graph except its effects, conditional on its direct causes.

Thus, we expect heart attacks to be independent of obesity conditional on diabetes and the change in the barometer reading to be independent of storm conditional on atmospheric pressure. The CMC is a generalisation of the screening-off condition found in various probabilistic theories of causation. It provides the link from correlation to causation: if there is a correlation, then it must have a causal explanation.

For causal inference, another condition is essential, *viz*. the Faithfulness condition (FC), according to which FC is a causal graph has only those probabilistic independencies that are entailed by the CMC.

FC says that all variables that are causally related are also correlated. It thus provides the link between correlation and causation in the reverse direction, from causation to correlation: if two variables are causally related, they are also correlated. That this does not always have to be the case is illustrated by Figure 29.2 (this is the above mentioned example due to Hesslow, 1976).

In the example Birth control affects Thrombosis via two independent routes: directly and indirectly through Pregnancy. If the causal strength of the two routes is identical and therefore the positive direct influence is exactly cancelled by the negative indirect influence, Birth control can be causally related, albeit not correlated with thrombosis.

The CMC, too, is subject to counterexamples. Not all correlations have a causal explanation, especially correlations among social science variables. In many cases, correlations can be induced by certain statistical properties of the time series describing variables, for instance when they are non-stationary (Sober, 1987, 2001; Hoover, 2003; Reiss, 2007). Moreover, common causes do not always screen off their joint effects – for example when causes operate genuinely indeterministically (Cartwright, 1999).

If both CMC and FC can be assumed to hold, simple algorithms can be applied to infer causal relations from statistics. According to the FC, causation entails correlation so all variables that are found to be correlated must be causally connected in one way or another. The CMC can then be used to distinguish direct, indirect and common causal relations. The theory of Bayes' nets provides a variety of algorithms that do precisely that. However, the basic idea is the foundation for all causal inference from statistics.

Process Tracing

Statistical methods of causal inference work only under fairly stringent conditions.

To begin with, they assume a tight link between causation and correlation, an assumption that is not always well warranted in social science research. But even if this very fundamental assumption can be made, statistical methods often fail for purely practical reasons. For example, one can infer causation from correlation – at best – when all common causes of the variables considered are measured. The complexity of the social world often makes this an insurmountable hurdle. Further, the statistical inference from observed frequencies to probabilities is only reliable when the samples are relatively large, and large sample size isn't always guaranteed in social research.

In such cases process tracing may be a viable alternative method of causal inference. Frequently, when a macro social variable causes another, its action is mediated by a more or less continuous process or mechanism. If this is so, there is uncertainty as to whether two specific variables are indeed causally related and social mechanisms are epistemically more readily accessible than relations between macro variables, knowledge about the mediating mechanism can be used for causal inference. In these cases, information about mechanisms provides evidence for hypotheses about causal relations among macro social variables of interest.

Sometimes the stronger claim that *only* knowledge about mechanisms provides sufficient evidence to provide a good reason to believe a causal hypothesis is made (Friedman and Schwartz, 1963: 59):

> However consistent may be the relation between monetary change and economic change, and however strong the evidence for the autonomy of the monetary changes, we shall not be persuaded that the monetary changes are the source of the economic changes unless we can specify in some detail the mechanism that connects the one with the other.

Necessary for successful causal inference or not, there is no doubt that if the mechanism is epistemically more readily accessible than the macro causal relation, learning about a mechanism connecting two social variables can be a powerful source of evidence for causal claims.

Sociologists call this method of inference 'process tracing'. Daniel Steel describes the method as follows (Steel, 2004: 67): 'Process tracing consists in presenting evidence for the existence of several prevalent social practices that, when linked together produce a chain of causation from one variable to another.' In the example discussed by Steel, the hypothesis at stake is Malinowski's (1935) claim that the possession of many wives was a cause of wealth and influence among Trobriand chiefs. The social practises that constitute evidence for the hypothesis are (a) the custom whereby brothers contribute substantial gifts of yams to the households of their married sisters; and (b) the fact that political endeavours and public projects undertaken by chiefs are financed primarily with yams.

While certainly a useful alternative method of causal inference, process tracing too has serious limitations. And obvious one is that the 'facts' used for process tracing such as those reported in (a) and (b) above have to be substantiated with evidence themselves. Together they are supposed to form a causal chain, so each link is itself causal in nature and must be substantiated with adequate methods. While often at this level other tools are available – for instance, participant observation as in the Malinowski case – there is no guarantee that the problem of inferring causal relations from observations is more easily solvable than at the aggregate or social level. And even in the best case, the results of process tracing are pretty modest. The method can only be used to establish a purely qualitative claim about the causal connection between two variables. It may well be the case that there are other mechanisms that undermine the effect of the process studied. For instance, the correctness of Malinowski's reasoning is not inconsistent with the possession of many wives actually being a prohibitor of wealth and influence because of other mechanisms that are quantitatively

stronger than the one established. Moreover, many disagreements in the social sciences are about the quantitative strength of a cause, not whether or not one variable is linked to another 'in one way or another'.

Policy Inference

Whereas descriptive and explanatory inferences are past- and present-regarding, policy inferences concern the future. Social scientists do not only aim to describe and explain: they also try to anticipate future events to facilitate planning, and to prepare policy decisions. In this final section of the overview of methods of evidence generation, I will look at two instances of sources of evidence for policy: expert political judgement and evidence-based policy.

Expert Political Judgement

Every day, countless 'experts' make predictions about political events that may or may not occur some time in the future. The ability to foresee such events, be they the outcome of a national election, the outbreak of a war or the end of a political era, with a reasonable degree of accuracy would be enormously useful for political decision makers, investors and society at large, if it could be achieved.

But sceptics argue that successful prediction is unattainable. There are two types of sceptics: those who deny that the world is predictable and those who deny that humans have the cognitive capacities to make successful predictions. Economic historian and methodologist Deirdre McCloskey belongs to the former group. She argues essentially that successful predictions are self-defeating because people would try to capitalise on and there by undermine them (McCloskey, 1998: 150–151). Philosophers Michael Bishop and J.D. Trout belong to the latter group. They argue that cognitive limitations such as memory and computing deficiencies as well as psychological phenomena such as overconfidence prevent us from achieving predictive success to the extent that the nature of the world would allow it (Bishop and Trout, 2005).

A recently published report of over 20 years of research on political experts by Philip Tetlock strikes a balance between these sceptics and meliorists – those who maintain that 'the quest for predictors of good judgment, and ways to improve ourselves, is not quixotic and that there are better and worse ways of thinking that translate into better and worse judgements [regarding future political events]' (Tetlock, 2006: 19). Many of his results confirm the sceptic (Tetlock, 2006: Chapter 2). For example, human experts' subjective probability judgements of outcomes are no better calibrated to the frequencies of these outcomes than those of a dart-throwing (randomising) chimpanzee who assigns an equal probability to all outcomes.[11] Whether the forecaster was an expert in the relevant field made hardly any difference to overall predictive accuracy, nor did education, experience, gender or political orientation. Moreover, statistical models of various degrees of sophistication beat even the best expert.

On the other hand, certain factors did make a difference to some outcomes and relative to some baselines. For instance, while randomising achieves a higher score on calibration, humans beat chimps with respect to discriminating between high and low probability events.[12] Experts perform better than undergraduates on both calibration as well as discrimination scores. Short-term forecasts are more accurate than long-term forecasts.

By far the most informative factor about an expert's judgement is his cognitive style. Tetlock uses Isaiah Berlin's metaphor of hedgehogs and foxes to characterise experts' cognitive style. Hedgehogs are thinkers who know one big thing and try to systematise every fact within the explanatory schema of that one big thing. Foxes know many little things, are sceptical of grand schemes and excel in 'ad hocery'. Foxes outperform hedgehogs on both calibration as well as discrimination and the best come close to some statistical models. Controlling for cognitive

style also changes the interpretation of other effects. Thus, while expertise has no across-the-bard effect, it is beneficial for foxes and outright harmful for hedgehogs. Foxes also score higher in the long-term than in the short-term while the opposite is true of hedgehogs (Tetlock, 2006: Chapter 3).

The important lesson from Tetlock's study is that there are better and worse experts, and there are ways to tell who is what. Tetlock's results confirm those reported by Bishop and Trout in that even the foxiest experts are outperformed by statistical models. But to the extent that political expertise is likely to maintain an important role in society, we had better know who to trust, how far and with respect to what claims.

Evidence-Based Policy

The evidence-based practise[13] movement can be understood as a reaction to what was perceived as over-reliance on expertise and related sources of knowledge such as folklore and tradition. These, in the eyes of proponents of the movement, unreliable guides to practise should be substituted by rigorously established scientific evidence, and many regard the randomised controlled trial (RCT) as the 'gold standard' of evidence. The movement became prominent first in medicine and other fields of health care and policy but is now gaining popularity in social and public policy in the United States and many other countries. RCTs have been conducted to study questions as diverse as the effect of CCTV surveillance on crime, class size on academic achievement, cognitive-behavioural treatment on anti-social behaviour, correctional boot camps on offending and many more.

In a RCT eligible subjects are divided into two groups using a random number generator. The aim of randomisation is to create two groups that are 'exchangeable' from a statistical point of view; that is, identical in all respects relevant for the assessment of the treatment effect. One group is assigned a treatment while the other figures as control group (and either remains untreated, receives an alternative treatment or a placebo). In a double-blind trial neither the participating subjects nor the treatment administrators know which is the treatment and which the control group. There are also multiple-blind trials in which also the statistician who analyses the data and other researchers are blinded.

RCTs are regarded as the gold standard of evidence in evidence-based practise because they are, if implemented successfully, a highly reliable sources of evidence for causal claims. But there are two catches. The first is indicated by the qualification 'if implemented successfully': RCT results are certain only under highly stringent, and indeed unrealistic conditions. These conditions include that the set of all other factors that affect the outcome are distributed identically between the two groups and that correlations always have a causal explanation (see for instance Cartwright, 2007a). However, randomisation in no way guarantees that treatment and control group are identical with respect to all confounders (see for instance Worrall, 2002), and especially in social-science applications, correlations have a variety of non-causal sources (Reiss, 2007).

The second catch is that the result of a RCT, even if it was implemented successfully, while known with certainty, is of very limited use. What a RCT at best proves is that a treatment has a causal effect *on average* and *in the population studied*. If there is an average causal effect, then the treatment must be effective in at least some individuals but we don't know which. In particular, it is not inconsistent that a treatment should be beneficial on average and yet harmful for some individuals. Further, the RCT result is valid only with respect to the particular arrangement of confounders present in the experiment. It is not informative about the effectiveness of the treatment in populations with a different arrangement of confounding factors.

The latter difficulty has come to be called the 'problem of external validity': if a test result is valid for an experimental population,

how do we apply it outside the experiment, 'in the field'? This issue has been taken up in recent philosophy of science and now there exist a variety approaches to deal with it: based on knowledge of mechanisms (Steel, 2008), on analogy (Guala, forthcoming, 2005) and on causal capacities (Cartwright, 2009). But none of these have the logical stringency of the ideal RCT. Therefore, the certainty associated with testing a policy proposition using RCTs is at least to some extent illusory.

Integrating, Weighing and Aggregating Evidence

A serious problem arises when pieces of evidence tell different stories about the hypothesis at stake. This can happen when, for instance, estimated correlations vary greatly between different studies or when different sources of evidence (e.g. statistical versus mechanistic evidence) give incompatible results. How do we combine such conflicting items of evidence in a way such as to draw reliable inferences regarding a hypothesis? Peter Achinstein discusses the following strategies (Achinstein, 2001: 124):

> Write a review article that summarises the different studies and results, without attempting to resolve the issue.
>
> 1 Choose a single, favourite study from the set and agree with its conclusions.
> 2 Compute overall averages for relevant statistics across the entire set of studies, independently of the sizes of the sample in each study or the conditions under which the samples were taken.
> 3 Take a vote. If a majority of the studies favour one conclusion, then that is the conclusion supported by the studies.
> 4 Employ meta-analysis, which its proponents regard as a much more sophisticated and reliable set of methods than 3 and 4 above.

(1) Obviously doesn't solve the problem. (2) seems arbitrary unless the chosen study is the only one without clearly identifiable methodological flaws. If either several studies are not subject to methodological problems or all studies have some flaws, the question simply reappears. (3) is a less sophisticated version of (5), which will be dealt with in more detail below. (4) again seems arbitrary. What if the best studies support a conclusion different from the majority? This strategy also works at best for simple yes/no results such as whether or not a treatment is effective. Many disagreements, however, are about the size of the effect, not its presence or absence.

Meta-analyses combine research results from a range of studies in a quantitative way. Many meta-analyses identify a common metric of effect size and model it using some form of regression in which the results of the individual studies figure as inputs. Meta-analyses have a variety of advantages over alternatives, including an increase in statistical power (over individual studies) and the ability to control for a variety of sources of error (see Hunter and Schmidt, 2004: Chapter 2). But they also come with serious drawbacks and limitations.

There are two obvious limitations. The first is that meta-analyses can only combine *statistical* evidence. The important problem of integrating evidence from different sources is not addressed. Second, the method requires that the individual studies deal with the same hypothesis. It is frequently not clear, however, what that amounts to.

But even if one restricts coverage to statistical evidence and is able to formulate reasonable inclusion criteria, meta-analysis is subject to criticisms. The most important of these is that the method assumes that differences between the studies are due to statistical error alone whereas in fact they often arise systematically. If, say, one study shows a positive treatment effect and another shows a negative effect or none, this may be due to differences in the causal structures characterising the two populations. Averaging over the two studies masks these differences and treats the samples as if drawn from one underlying population. This way important information is lost. Similar considerations apply to the choice of measurements used and other aspects of the study design.

In sum, to the extent that individual studies are biased for statistical reasons (for instance, because of sampling error), meta-analyses are a powerful tool to reduce these types of bias. They do not, however, eliminate the need for context-sensitive judgements about the quality of the individual studies entering the analysis.

EVIDENCE FOR USE

All scientific methods are associated with specific types of hypotheses the researcher is entitled to infer by the evidence generated by them. Often, there is a kind of trade-off: the more reliable the method (that is, the more secure the inference based on the evidence produced by it), the narrower the range of hypotheses that can be supported by that method. Some methods, such as the RCT, 'clinch' their results: under a suitable set of background assumptions, the evidence deductively entails the hypothesis. But the epistemic certainty is bought at a cost: the situations of which the assumptions are likely to be true are very rare. Other methods, such as participant observation only 'vouch' for their results: the evidence makes the hypothesis more likely without proving it. These methods tend to be more broadly applicable but there always remain reasons to believe that the conclusion is false (for the clincher/voucher dichotomy, see Cartwright, 2007b).

The value of knowing a hypothesis is constrained by the certainty with which it is known but also by the value of the ways in which it can be put to use. Philosophers in the analytic tradition have tended to be impressed by the paradigm of 'exact science' and consequently focused on the former at the expense of the latter. But for many important applications, there simply are no techniques that produce evidence of the desired quality.

Cartwright (2006) argues that philosophers, rather than develop methods with certain epistemologically desirable characteristics, should instead start from scientifically – and politically – desired uses and work on methods designed to target hypotheses knowledge of which is advantageous in their light. At the core of her project is a kind of contextualism regarding evidence (Cartwright, 2006: 983): 'What justifies a claim depends on what we are going to do with that claim, and evidence for one use may provide no support for others'. 'Evidence for use' is a research project that urges philosophers and methodologists to pay closer attention to the demands of scientific practise and social utility. Arguably, the foundational debates in the social sciences can profit from such a reorientation too. Much philosophical effort is spent, perhaps needlessly, on debating whether social science is real science, whether one should or shouldn't study the social world with essentially the same methods as the natural world, how to effectively separate 'facts' and 'values', whether there is such a thing as society or rather a mere heap of individuals – while significant social issues remain unnoticed by the philosophers claiming to be experts in these debates.

NOTES

1 These three inferences allude to enumerative induction, Bayesianism and the error-statistical approach, respectively. See below for detailed descriptions.

2 If f and φ are exchanged, the same method is used to argue that f is an effect of φ.

3 Once more, if f and φ are exchanged, the same method is used to argue that f is an effect of φ.

4 This interpretation was at the heart of Carnap's 1950 influential theory of evidence.

5 Mayo inserts a footnote here saying that she prefers to phrase this in terms of data e being a 'good indication' of H.

6 See also Hon (1998) and Carrier (2001) for similar criticisms. It should be noted, however, that error statisticians *claim* to have a philosophy of evidence and induction of entirely general scope: see Mayo (1997, 2000, 2004); Mayo and Spanos (2004, 2006).

7 Achinstein first defines only a necessary condition but later qualifies: 'the conditions in (PE) are proposed as both necessary and sufficient' (2001).

8 'Being male' might explain Wheaties eating behaviour in the same sense that gender explains other preferences. Achinstein assumes that this isn't the case here.

9 Though many Bayesians require satisfaction of the so-called 'principal principle', which says that if an agent knows the physical probability of an outcome, his degree of belief should be the same. The term is due to David Lewis (1980).

10 Some authors, in the methodological literature most notably Kevin Hoover (2003) distinguish the two notions. With most other philosophers I will glance over the differences here (for a discussion, see Reiss, 2007).

11 What is meant by 'calibrated' here is that the subjective probability judgements reflect the objective frequencies of those types of events. For example, those events assigned a probability of 10 per cent should actually happen in 10 per cent of the cases.

12 Perfect discrimination is achieved when all those events that obtain are predicted as 'certain' and all those that do not obtain as 'impossible'.

13 With 'evidence-based practise' I refer to evidence-based movements in all branches of knowledge creation and policy such as medicine, health care and policy as well as management, social and public policy. 'Evidence-based policy' is narrower, covering only the latter two fields. To my knowledge, there is no standardised terminology in this area.

REFERENCES

Achinstein, P. (2001) *The Book of Evidence*. Oxford: Oxford University Press.

Bird, A. (2007) 'Inference to the only explanation', *Philosophy and Phenomenological Research*, 74(2): 424–432.

Bishop, M. and J.D. Trout (2005) *Epistemology and the Psychology of Human Judgment*. Oxford: Oxford University Press.

Carnap, R. (1950) *Logical Foundations of Probability*. Chicago, IL: University of Chicago Press.

Carrier, M. (2001) 'Critical notice: Error and the growth of experimental knowledge', *International Studies in the Philosophy of Science*, 15(1): 93–98.

Cartwright, N. (1999) *The Dappled World*. Cambridge: Cambridge University Press.

Cartwright, N. (2007a) 'Are RCTs the gold standard?', *BioSocieties*, 2(2): 11–20.

Cartwright, N. (2007b) *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

Cartwright, N. (2009) 'Evidence-based policy: What's to be done about relevance', *Philosophical Studies*, 143(1): 127–136.

DeWalt, K. and B. DeWalt (2002) *Participant Observation: A Guide for Fieldworkers*. Walnut Creek, CA: AltaMira Press.

Friedman, M. and A. Schwartz (1963) 'Money and business cycles', *Review of Economics and Statistics*, 45(1, Part 2, Supplement): 32–64.

Guala, F. (2005) *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Guala, F. (forthcoming) 'Extrapolation Without Process Tracing', *Philosophy of Science*, PSA 2008.

Hacking, I. (1965) *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hacking, I. (1972) 'Likelihood', *British Journal for the Philosophy of Science*, 23: 132–137.

Harman, G. (1965) 'Inference to the best explanation', *Philosophical Review*, 74(1): 88–95.

Hempel, C. (1945) 'Studies in the logic of confirmation (I.)', *Mind*, 54(213): 1–26.

Hempel, C. (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York, NY: Free Press.

Hesslow, G. (1976) 'Discussion: Two notes on the probabilistic approach to causality', *Philosophy of Science*, 43: 290–292.

Hon, G. (1989) 'Towards a typology of experimental errors: An epistemological view', *Studies in History and Philosophy of Science*, 20: 469–504.

Hon, G. (1995) 'Is the identiciation of experimental error contextually dependent? The case of Kaufmann's experiment and its varied reception', in J. Buchwald (ed.) *Scientific Practice: Theories and Stories of Doing Physics*. Chicago, IL: University of Chicago Press. pp. 170–223.

Hon, G. (1998) 'Exploiting errors', *International Studies in the Philosophy of Science*, 29(3): 465–479.

Hoover, K. (2003) 'Nonstationary time-series, cointegration, and the principle of the common cause', *British Journal for the Philosophy of Science*, 54: 527–551.

Howson, C. and P. Urbach (1993) *Scientific Reasoning: The Bayesian Approach*. 2nd edn. Chicago, IL: Open Court.

Hunter, J. and F. Schmidt (2004) *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: Sage.

Kincaid, H. (2007) 'Contextualist morals and science', in H. Kincaid, J. Dupré and A. Wylie (eds) *Value-Free Science? Ideals and Illusions*. Oxford: Oxford University Press. pp. 218–238.

Klein, J. (2003) 'Francis Bacon', *Stanford Encyclopedia of Philosophy (Spring 2009)* E. Zalta (ed.), URL = < http://plato.stanford.edu/archives/spr2009/entries/francis-bacon/> (accessed 23 November 2009) .

Lewis, D. (1980) 'A subjectivist's guide to objective chance', R. Jeffrey (ed.) *Studies in Inductive Logic*

and Probability. Berkeley and Los Angeles: University of California Press. pp. II.

Lipton, P. (2004) *Inference to the Best Explanation*. 2nd edn. London: Routledge.

Mackie, J. (1974) *The Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press.

Malinowski, B. (1935) *Coral Gardens and Their Magic*. New York, NY: American Book Co.

Mayo, D. (1996) *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, D. (1997) 'Error statistics and learning from error: Making a virtue of necessity', *Philosophy of Science*, 64(PSA 1996): S195–212.

Mayo, D. (2000) 'Experimental practice and an error statistical account of evidence', *Philosophy of Science*, 67(Proceedings): S193–207.

Mayo, D. (2004) 'An error-statistical philosophy of evidence', in M. Taper and S. Lele (eds) *The Nature of Scientific Evidence*. Chicago, IL: University of Chicago Press. pp. 79–96.

Mayo, D. and A. Spanos (2004) 'Methodology in practice: Statistical misspecification testing', *Philosophy of Science*, 71: 1007–1025.

Mayo, D. and A. Spanos (2006) 'Severe testing as a basic concept in a Neyman–Pearson philosophy of induction', *British Journal for the Philosophy of Science*, 57: 323–357.

McCloskey, D. (1998) *The Rhetoric of Economics*. 2nd edn. Madison, WN: University of Wisconsin Press.

Mill, J. S. (1843 [1874]) *A System of Logic*. New York, NY: Harper.

Norton, J. (2003) 'A material theory of induction', *Philosophy of Science*, 70(4): 647–670.

Norton, J. (2008) 'Must evidence underdetermine theory?', in M. Carrier, D. Howard and J. Kourany (eds) *The Challenge of the Social and the Pressure of Practice*. Pittsburgh, PA: Pittsburgh University Press. pp. 17–44.

Okasha, S. (2000) 'Van Fraassen's critique of inference to the best explanation', *Studies in the History and Philosophy of Science*, 34(4): 691–710.

Quine, W.v.O. (1969) 'Natural kinds', in W.v.O. Quine (ed.) *Ontological Relativity and Other Essays*. New York, NY: Columbia University Press. pp. 114–138.

Ragin, C. (1998) 'The logic of quality comparative analysis', *International Review of Social History*, 43(Supplement): 105–124.

Reiss, J. (2007) 'Time series, nonsense correlations and the principle of the common cause', in F. Russo and J. Williamson (eds) *Causality and Probability in the Sciences*. London: College Publications. pp. 179-196.

Reiss, J. (2008) *Error in Economics: Towards a More Evidence-Based Methodology*. London: Routledge.

Salmon, W. (1975) 'Confirmation and relevance', G. Maxwell and R. Anderson (ed.) *Induction, Probability, and Confirmation*. Don Mills, ON: Burns & Maceachern. VI: 3–36.

Schickore, J. (2005) '"Through thousands of errors we reach the truth" – but how? On the epistemic roles of error in scientific practice', *Studies in History and Philosophy of Science*, 36: 539–556.

Scriven, M. (1966) 'Causes, connections and conditions in history', in W. Dray (ed.) *Philosophical Analysis and History*. New York, NY: Harper and Row. pp. 238–264.

Scriven, M. (2008) 'A summative evaluation of RCT methodology and an alternative approach to causal research', *Journal of MultiDisciplinary Evaluation*, 5(9): 11–24.

Sober, E. (1987[1994]) 'The principle of the common cause', in *From a Biological Point of View*. Cambridge: Cambridge University Press. pp. 158–174.

Sober, E. (2001) 'Venetian sea levels, British bread prices, and the principle of the common cause', *British Journal for the Philosophy of Science*, 52: 331–346.

Steel, D. (2004) 'Social mechanisms and causal inference', *Philosophy of the Social Sciences*, 34(1): 55–78.

Steel, D. (2005a) 'The facts of the matter: A discussion of Norton's material theory of induction', *Philosophy of Science*, 72: 188–197.

Steel, D. (2005b) 'Indeterminism and the causal Markov condition', *British Journal for the Philosophy of Science*, 56: 3–26.

Steel, D. (2008) *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.

Sugden, R. (2000) 'Credible worlds: The status of theoretical models in economics', *Journal of Economic Methodology*, 7(1): 1–31.

Tetlock, P. (2006) *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton: Princeton University Press.

van Fraassen, B. (1985) *Laws and Symmetry*. Oxford: Oxford University Press.

Worrall, J. (2002) 'What evidence in evidence-based medicine', *Philosophy of Science*, 69: S316–330.