To conclude, the analysis above suggests that both solutions above may be seen to conflate somehow. For, if *common cause completability* is applied to the 'Venice-Britain' scenario we obtain physically meaningless screening-off events. And this suggests that the 'Venice-Britain' correlation is just an artificial one. Thus, whether 'Venetian sea levels and British bread prices' does constitute indeed a genuine counterexample to RPCC is just a matter of whether or not *purely formal* correlations are in need of (causal) explanation.

References

Cartwright, N. 1987. How to tell a common cause: Generalizations of the conjunctive fork criterion. In *Probability and causality*, ed. J.H. Fetzer, 181–188. Dordrecht: Reidel.

Hofer-Szabó, G., M. Rédei, and L.E. Szabó. 1999. On Reichenbach's common cause principle and reichenbach's notion of common cause. *The British Journal for the Philsophy of Science* 50: 377–399.

Hofer-Szabó, G., M. Rédei, and L.E. Szabó. 2000a. Common cause completability of classical and quantum probability spaces. *International Journal of Theoretical Physics* 39: 913–919.

Hofer-Szabó, G., M. Rédei, and L.E. Szabó. 2000b. Reichenbach's common cause principle: Recent results and open questions. *Reports on Philosophy* 20: 85–107.

Hoover, K.D. 2003. Nonstataionary time series, cointegration, and the principle of the common cause. *The British Journal for the Philosophy of Science* 54: 527–551.

Rédei, M. 2002. Reichenbach's common cause principle and quantum correlations. In *Non-locality, modality and Bell's theorem*, eds. T. Placek and J. Butterfield, 259–270. Dordrecht: Kluwer.

Reichenbach, H. 1956. *The direction of time; Edited by Maria Reichenbach*. Unabridged Dover (1999) republication of the original University of California Press (1956) publication.

Salmon, W. 1984. *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

San Pedro, I., and M. Suárez. 2009. Reichenbach's principle of the common cause and indeterminism: A review. In *Philosophical essays in physics and biology*, ed. J.L. González Recio, 223–250. Hildesheim, Zürich, New York: Georg Olms.

Sober, E. 1987. The principle of the common cause. In *Probability and causality: Essays in honor of Wesley Salmon*, ed. J.H. Fetzer, 211–228. Dordrecht: Reidel.

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

Steel, D. 2003. Making time stand still: A response to Sober's counter-example to the principle of the common cause. *The British Journal for the Philosophy of Science* 54: 309–317.

Yule, G.U. 1926. Why do we sometimes get nonsensical relations between time series? A study of sampling and the nature of time series. *Journal of the Royal Statistical Society* 89: 1–64.

Chapter 30
Coincidences and How to Reason About Them

Elliott Sober

30.1 A Familiar Dialectic

The naïve see causal connections everywhere. Consider the fact that Evelyn Marie Adams won the New Jersey lottery twice. The naïve find it irresistible to think that this cannot be a coincidence. Maybe the lottery was rigged or perhaps some uncanny higher power placed its hand upon her brow. Sophisticates respond with an indulgent smile and ask the naïve to view Adams' double win within a larger perspective. Given all the lotteries there have been, it isn't at all surprising that someone would win one of them twice. No need to invent conspiracy theories or invoke the paranormal—the double win was a mere coincidence.

The naïve focus on a detailed description of the event they think needs to be explained. The *New York Times* reported Adams' good fortune and said that the odds of this happening by chance are 1 in 17 trillion; this is the probability that Adams would win both lotteries if she purchased a single ticket for each and the drawings were at random. In fact, the newspaper made a small mistake here. If the goal is to calculate the probability of Adams' winning those two lotteries, the reporter should have taken into account the fact that Adams purchased multiple tickets; the newspaper's very low figure should have been somewhat higher. However, the sophisticated response is that this modest correction misses the point. For sophisticates, the relevant event to consider is not that Adams won those two lotteries, but the fact that *someone* won two state lotteries at *some* time or other. Given the many millions of people who have purchased lottery tickets, this is "practically a sure thing" (Diaconis and Mosteller 1989, Myers 2002).

Another example of reasoning about coincidence in which the same dialectic unfolds begins with the fact that my birthday (06061948) occurs at the 16,769,633th

E. Sober (✉)
Department of Philosophy, University of Wisconsin – Madison, Madison, WI, USA
e-mail: ersober@wisc.edu

position of the decimal expansion of π (not counting the initial "3").[1] The probability of this occurring is very small, if numbers appear at random in the decimal expansion. The naïve conclude that my birthday's occurring at that exact position cannot be a mere coincidence; perhaps my date of birth was so arranged that the number 16,769,633 would provide me with an encrypted message that points the way to my destiny. The sophisticated reply that the probability of my birthday's occurring somewhere in the first 100 million digits is actually very high—about 2/3. Given this, there is no reason to think that my birth date's showing up where it does is anything but a coincidence.

30.2 How the Naive and the Sophisticated Reason

The naïve and the sophisticated[2] agree about one thing but disagree about another. Both rely on a rule of inference that I will call *probabilistic modus tollens*. This is the idea that you should reject a hypothesis if it tells you that what you observe is enormously improbable. The naïve think that the hypothesis of Mere Coincidence strains our credulity too much. Since the hypothesis of Mere Coincidence says that the probability of Adams' double win is tiny, we should reject that hypothesis. Sophisticates grant the authority of probabilistic *modus tollens*, but contend that the hypothesis of Mere Coincidence should be evaluated by seeing what it says about the observation that someone wins two state lotteries at some time or other. Since this is very probable according to the hypothesis of Mere Coincidence, we should decline to reject that hypothesis. The naïve and the sophisticated thus seem to agree on the correctness of *probabilistic modus tollens*. Their disagreement concerns how the event to be explained should be described.

Sophisticates avoid rejecting the hypothesis of Mere Coincidence by replacing a logically stronger description of the observations with one that is logically weaker. The statement

(1) Evelyn Adams, having bought four tickets for each of two New Jersey lotteries, wins both.

is logically stronger than the statement

(2) Someone at sometime, having bought some number of tickets for two or more lotteries in one or more states, wins at least two lotteries in a single state.

It is a theorem in probability theory that logically weakening a statement can't lower its probability—the probability will either go up or stay the same. In the case at

[1] Go to http://www.angio.net/pi/piquery to see if your birthday appears in the first 100 million digits.

[2] The naïve and the sophisticated are characters in my story; I do not mean to suggest that all sophisticated thinkers in the real world reason exactly in the way I'll describe the sophisticated as reasoning.

hand, the hypothesis of Mere Coincidence says that (1) is very improbable, but that (2) is very probable.

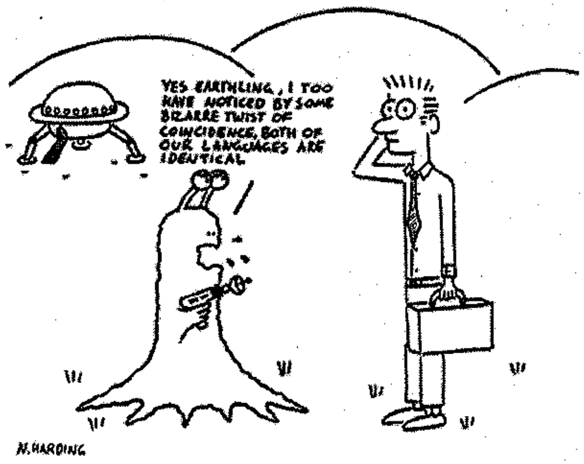
Diaconis and Mosteller (1989, 859) say that the relevant principle to use when reasoning about coincidences is an idea they term the *Law of Truly Large Numbers*. This says that "with a large enough sample, any outrageous thing is likely to happen." They cite Littlewood (1953) as having the same thought; with tongue in cheek, Littlewood defined a miracle as an event whose probability is less than 1 in a million. Using as an example the U.S. population of 250 million people, Diaconis and Mosteller observe that if a miracle "happens to one person in a million each day, then we expect 250 occurrences a day and close to 100,000 such occurrences a year." If the human population of the earth is used as the reference class, miracles can be expected to be even more plentiful.

30.3 Two Problems for Sophisticates

Sophisticates bent on using probabilistic *modus tollens* should be wary about the strategy of replacing a logically stronger description of the observations with one that is logically weaker. The reason for wariness is that this strategy allows one to decline to reject hypotheses of Mere Coincidence no matter what they are and no matter what the data say. Even when there is compelling evidence that the observations should *not* be explained by this hypothesis, the hypothesis of Mere Coincidence can be defended by logically weakening the observations..

Consider, for example, Alfred Wegener's (1924) defense of the hypothesis of continental drift. Wegener noticed that the wiggles in the east coast of South America correspond rather exactly to the wiggles in the west coast of Africa. The pattern is "as if we were to refit the torn pieces of a newspaper by matching their edges and then check whether the lines of print run smoothly across (Wegener 1924, 77)." Wegener also noticed that the distribution of geological strata down one coast matches the distribution down the other. In addition, he observed that the distribution of organisms down the two coasts—both fossilized and extant—shows the same detailed correlation. Wegener argued that this systematic matching should not be explained by the hypothesis of Mere Coincidence. His preferred alternative was that the continents had once been in contact and then had drifted apart.

Wegener encountered intense opposition from geophysicists, who didn't see how continents could plough through the ocean floor. I will return to this criticism later. My present point is that it would have been bizarre to counter Wegener's argument by weakening the data. A sophisticate bent on retaining the hypothesis of Mere Coincidence could point out that there are billions of planets in the universe that contain continents separated by wide oceans. If wiggles in coast lines and distributions of geological strata and of organisms are in each continent independently caused, there surely will exist at least one pair of continents on some planet or other that exhibits the kind of matching that Wegener found so interesting. With the data suitably weakened, probabilistic *modus tollens* no longer tells you to reject the hypothesis of Mere Coincidence.



A similar point is illustrated by the accompanying cartoon (reprinted here with the kind permission of its creator, Nick Harding). If life forms from another planet turn out to speak English, the irresistible inference will be that we and they have had some sort of prior contact. The idea that the detailed resemblance of the two languages is a Mere Coincidence strains our credulity too much. However, if we wish to hold fast to the belief that the resemblance is a Mere Coincidence, we can avoid having probabilistic *modus tollens* force us to reject that hypothesis merely by weakening our description of what the two languages have in common. Instead of focusing on the fact that the two languages match in a thousand specific ways, we can restrict our attention to the modest fact that both contain nouns. We then can reply that it isn't at all surprising that two languages should both contain nouns if they developed independently; after all, nouns are useful.[3] Notice that I just weakened the description of the data in a way that differs from the kind of weakening I considered in connection with Wegener. I didn't ask what the probability is that somewhere in the universe two languages would match even though they evolved independently (which is not to deny that that question will lead to the same conclusion). This brings out a further problem with the strategy of weakening the data at will. There are many ways to weaken the data. Which weakening should one employ? Why not simply replace the data with a tautology?

I began by noting that the naïve seem to think that *nothing* is a Mere Coincidence. Sophisticates who constantly weaken their description of the data to avoid rejecting hypotheses of Mere Coincidence seem to think that *everything* is a Mere Coincidence. These sophisticates are not just sophisticated—they are *jaded*. No correlation, no matter how elaborate and detailed, impresses them. In fact, none can

[3] Darwin (1859, ch. 13) argued that adaptive similarities between species provide poor evidence for common ancestry and that it is useless and deleterious similarities that provide more powerful evidence; see Sober (2008, 2011) for discussion. Darwin (1871, ch. 6) noticed the parallel epistemological problems that connect historical linguistics and phylogenetic inference.

impress them; their trick of weakening the data works against all comers. What we need is guidance on when the description of the data may be weakened, not the imperative to always do so or the permission to do so whenever we please.

Statistics provides guidance on the question of when one's description of the data may be weakened. It is given in the theory of *sufficient statistics*. R.A. Fisher (1959) introduced this idea in the context of his theory of point estimation. Suppose you want to estimate a coin's probability θ of landing heads when tossed and you assume that the tosses are independent and identically distributed (i.i.d.) – each toss has the same probability of landing heads and the results on some tosses don't influence the probabilities of others. To figure out which estimate is best, you toss the coin 1000 times, say, and obtain a particular sequence of heads and tails. Do you need to use this exact sequence as your description of the data, or can you just attend to the number of heads (which, let us suppose, was 503)? As it happens, this weaker description *suffices*; it is a sufficient statistic in the sense that it captures all the evidentially relevant information that the exact sequence contains. More specifically, the frequency of heads is a sufficient statistic in the context of using *maximum likelihood estimation* (MLE) as one's method for estimating θ because

(3) $$\frac{\Pr(\text{the exact sequence}|\theta = p)}{\Pr(\text{the exact sequence}|\theta = q)} = \frac{\Pr(\text{the number of heads}|\theta = p)}{\Pr(\text{the number of heads}|\theta = q)}$$

In all these conditional probabilities, I assume that the coin was tossed 1000 times. The reason (3) is true is that

(4) $$\Pr(\text{the exact sequence}|\theta = x) = x^{503}\,(1-x)^{497}$$

and

(5) $$\Pr(\text{number of heads}|\theta = x) = \binom{1000}{503} x^{503}\,(1-x)^{497}$$

This is why the left-hand and right-hand ratios in (3) must have the same value. The maximum likelihood estimate of θ is the same whether you use the stronger or the weaker description of the data, and the likelihood ratio of that best estimate, compared to any inferior estimate, will be the same, again regardless of which description of the data you use. Notice that what counts as a sufficient statistic depends on the method of inference you use and on the range of possible hypotheses you want to consider.[4] In the example just described, MLE is the method used and the assumption is that tosses are i.i.d. If MLE were used in the context of testing whether tosses are independent of each other, the number of heads would not be a sufficient statistic; information about the exact sequence would additionally be relevant.

[4] Notice also that the argument that appeals to (3) to show that the number of heads is a sufficient statistic depends on using the likelihood *ratio* as the relevant method for comparing the two estimates. If the likelihood *difference* were used instead, the corresponding equality would not be true. How one measures weight of evidence matters; see Fitelson (1999) for further discussion.

With these ideas in mind, let's return to the example of Evelyn Marie Adams' double win in the New Jersey lottery. If we use probabilistic *modus tollens*, the weakened description of the data given in (2) is *not* endorsed by the idea of sufficient statistics. The point is that shifting from (1) to (2) makes a difference in the context of probabilistic *modus tollens*, even though shifting from (4) to (5) does not matter from the point of view of MLE under the i.i.d. assumption. Shifting from a highly specific description of the data to one that is logically weaker is often permissible, but that is not enough to justify the sophisticate's pattern of reasoning about Adams. If a statistic is sufficient, you are *permitted* to shift to that weaker description of the data; you are not *obliged* to do so. The shift is permissible when and only when it doesn't change what you infer.

The problem of whether to weaken one's description of the evidence, and how to do so, is a problem for the sophisticate, not for the naïve. However, there is a second problem that both must face—both rely on probabilistic *modus tollens*. This is a form of inference that no one should touch with a stick. The similarity between *modus tollens* and its probabilistic analog may suggest that the latter must be legitimate because the former is deductively valid; however, this is an illusion. *Modus tollens* says that if H entails O and O turns out to be false, then you should conclude that H is false. Probabilistic *modus tollens* says that if Pr(O | H) is very high and O turns out to be false, that you likewise should conclude that H is false. My beef with probabilistic *modus tollens* is not that the conclusion does not deductively follow from the premises. I've drawn a double line between premises and conclusion in Prob-MT below to acknowledge that this is so, but that isn't enough to rescue the principle. Rather, my objection is that the occurrence of an event that a hypothesis says is very improbable is often evidence *in favor* of the hypothesis, not evidence *against* it. What is evidence in favor of H cannot be a sufficient reason to reject H.

$$\text{(MT)}\quad \begin{array}{l} \text{If H then O.} \\ \text{not-O.} \\ \hline \text{not-H.} \end{array} \qquad\qquad \text{(Prob-MT)}\quad \begin{array}{l} \text{Pr(O|H) is very high.} \\ \text{not-O.} \\ \hline\hline \text{not-H.} \end{array}$$

Consider, for example, the use of DNA testing in forensic contexts. DNA evidence can be used to draw an inference about whether two individuals are related (for example, in paternity suits) or to draw an inference about whether a person suspected of a crime was at the crime scene. In both cases, you begin by determining whether two DNA samples match. This may seem to be a context in which probabilistic *modus tollens* is plausible. Suppose two individuals match at the genetic loci examined, and that the probability of this match is only, say, 6.5×10^{-38} if the two individuals are unrelated. This may seem to provide ample grounds for rejecting the hypothesis that the individuals are unrelated. However, what is missing from this exercise is any representation of how probable the data would be if the individuals *were* related. The National Commission on the Future of DNA Evidence report (2000, 66) discusses an example of this sort in which two individuals match at 13 loci for genes that happen to be rare. The authors of this report calculate the above figure of 6.5×10^{-38} as the probability of the data under the hypothesis that the individuals are unrelated. However, it also is true that if the individuals were full sibs, the probability of the match would be 7.7×10^{-32}. Surely it would be absurd to apply probabilistic *modus tollens* twice over, first rejecting the hypothesis that the two individuals are unrelated and then rejecting the hypothesis that they are related. In fact, the data lend support to the hypothesis that the two individuals are sibs; it would be wrong to use the data to reject that hypothesis. The evidence *favors* the hypothesis that the two individuals are sibs over the hypothesis that they are unrelated because the observations are more probable under the first hypothesis than they are under the second. This is the Law of Likelihood (Hacking 1965, Edwards 1972, Royall 1997, Sober 2008). It isn't the *absolute value* of the probability of the data under a single hypothesis that matters; rather, the relevant issue is how two such probabilities *compare*. The Law of Likelihood allows for the possibility that evidence may differentially support a hypothesis even though the hypothesis says that the evidence was very improbable. Notice also that the Law of Likelihood avoids an embarrassing question that defenders of probabilistic *modus tollens* must answer—how improbable is improbable enough for the hypothesis to be rejected? Defenders of Prob-MT have had to admit that this question has only a conventional answer.

What I have dubbed probabilistic *modus tollens* shows up in statistics in the form of Fisher's test of significance. There is more to a significance test than Prob-MT, but it remains true that significance tests reject a hypothesis (at a given "level of significance") when a certain description of the observations is less probable than a threshold that has been chosen arbitrarily. It is interesting to reflect on this practice in light of what Fisher (1959, 39) said about significance tests. According to Fisher, you have two choices when a hypothesis says that your observations are very improbable—either the hypothesis is false or something very improbable has just occurred. Fisher was right about the disjunction. However, what does not follow is that the hypothesis is false or that it is probably false; in fact, as just noted, it doesn't even follow that you have obtained evidence against the hypothesis (Hacking 1965, Edwards 1972, Royall 1997).

When the naïve and the sophisticated reason about whether Evelyn Marie Adams' double win was a Mere Coincidence, both help themselves to probabilistic *modus tollens*. Our task in what follows is to understand this problem without appealing to that faulty rule of inference. Sophisticates also allow themselves to violate the Principle of Total Evidence. They are happy to substitute a weaker description of the data for a stronger one, even though that changes the conclusion that the rule of inference they use instructs them to draw. We need to explain why the naïve are wrong to think that nothing is a Mere Coincidence without violating that principle. This may seem to return us to square one, but it does not. There is something right about the sophisticate's demand that the data about Evelyn Adams be placed in a wider perspective. We need to consider not just her double win, but the track records that others have had and whether she bought tickets in other lotteries that did not turn out to be winners. However, moving to this wider data set does not involve *weakening* the initial description of the data, but *adding* to it; the key is to make the data *stronger*.

30.4 Coinciding Observations, Coincidence Explanations, and Reichenbach's Principle of the Common Cause

Some regimentation of vocabulary is in order. First of all, what is a coincidence? Diaconis and Mosteller (1989, 853) suggest a working definition: a coincidence is "a surprising concurrence of events, perceived as meaningfully related, with no apparent causal connection." This is a good start, but it has the drawback of entailing that whether something is a coincidence is a subjective matter. There are two elements in this definition that we should separate. First, there is the idea of *coinciding observations*. When you and I meet on a street corner, our locations coincide. The same is true of the east coast of South American and the west coast of Africa—their wiggles, geological strata, and biogeography coincide. And perhaps it doesn't offend too much against the rules of English usage to say that the person who won the New Jersey lottery in one week "coincides" with the person who won it a few weeks later (they are identical). Observations coincide when they are similar in some respect. There is no need to be precise about how much (or what kind of) similarity is required for two observations to coincide, since the main point is to distinguish the observations from a kind of hypothesis that might be offered to explain them. Here we need the idea of a *coincidence explanation*. A coincidence explanation asserts that the observations are not causally connected. By this I mean that neither causes the other and they do not have a common cause. Thus, to say that it is a "mere" coincidence that two events are similar is to suggest a certain kind of explanation; each event was produced via a separate and independent causal process. Saying that the similarity of the observations is a coincidence does *not* mean that the similarity is inexplicable or that there is no need to explain the similarity. Understood in this way, it is an objective matter whether a given coincidence explanation is true, assuming as I will that causation is an objective matter.

With coinciding observations distinguished from coincidence explanations, we can kick away the ladder and see that coinciding observations are not required for the question to arise of whether a hypothesis of Causal Connectedness is superior to a hypothesis of Mere Coincidence. We sometimes need to consider this choice when the observations exhibit a pattern of *dis*similarity. Cartwright (1994, 117) describes the following example. Suppose I go shopping each week at a grocery store with $10 to spend. I spend some portion of the $10 on meat and the rest on vegetables. Suppose, when you observe my cash register receipts over the course of a year, you see that I never spend exactly $5 on the one and exactly $5 on the other. The dollar amounts never coincide. But the fact that they always sum to $10 is not a mere coincidence. They are two effects of a common cause. So observations need not be similar for the question of coincidence to arise. If you and I always order different desserts when we dine together at a restaurant, the waiter may rightly suspect that this is not a coincidence.

The match, or mismatch, of two *token* events is a rather small data set. When there are many pairs of token events, a pattern involving *kinds* of events may emerge. Based on the relative frequencies of kinds of events, one may infer that a correlation, either positive or negative, exists. Correlation is a probabilistic concept. Dichotomous event types A and B are positively correlated precisely when $Pr(A\&B) > Pr(A)Pr(B)$. Cartwright's shopping example involves a negative correlation; let A = my spending more than $5 on meat and B = my spending more than $5 on vegetables. If you infer the probabilities of these two event types from their frequencies in the data set describing my 52 trips to the grocery store, you'll infer that $Pr(A) \approx Pr(B) \approx \frac{1}{2}$, but that $Pr(A\&B) = 0$. Given a correlation (positive *or* negative), the question is whether the pattern of matching (or mismatching) of the token events should be explained by saying that the correlates are causally connected or by saying that the correlation is a mere coincidence.

Reichenbach (1956) elevated our natural preference for hypotheses of causal connection to the status of a metaphysical principle.[5] His *principle of the common cause* says that whenever two events are correlated, the explanation must be that the two correlates are causally connected. This principle is central to recent work on causal modeling and directed graphs (Spirtes et al. 2001, Pearl 2000, Woodward 2003). I think it is better to treat Reichenbach's idea as an epistemological principle that should be evaluated in terms of the Law of Likelihood (Sober 1988a, b, 2001, 2008). The question is whether a hypothesis of Causal Connection renders the observations more probable than does the hypothesis of Mere Coincidence. When this is so, the evidence favors the first hypothesis over the second; it does not guarantee that the Causal Connection hypothesis must be true.[6]

Reichenbach was able to show that the fact that two events are correlated deductively follows from a certain type of Common Cause model, one in which the postulated common cause raises the probability of each effect and renders them conditionally independent. Viewed from the point of view of the Law of Likelihood, Reichenbach's argument can be adapted to cases in which the *explanandum* is the coinciding of two token events, rather than the correlation of two event types (Sober 1988b). And the *mis*match of two events sometimes points towards a common cause explanation and away from a separate cause explanation, depending again on the details of how the common cause and separate cause hypotheses are formulated. Thus, in a wide range of cases, the question of whether it is a mere coincidence that the two events E_1 and E_2 occurred can be addressed by comparing the likelihood of the hypothesis of Causal Connection with the likelihood of the hypothesis of Mere Coincidence.

[5] I do not use the term "metaphysical" here in the pejorative sense sometimes used by logical positivists. Rather, my use of the term contrasts with "epistemological." The former has to do with the way the world is, the latter with the beliefs we should form about the world.

[6] One reason that Reichenbach's principle should not be formulated metaphysically is the fact that it is at least a defensible position to maintain that quantum mechanics describes event types that are lawfully correlated but not causally connected. Arthur Fine has pointed out to me that these correlations also show that my categories of Mere Coincidence and Causal Connection are not exhaustive.

30.5 The Limits of Likelihood

The Law of Likelihood is a useful tool in the project of reasoning about coincidences, but it doesn't provide the complete epistemology we need. The problem is that likelihood considerations favor hypotheses of causal connection in contexts in which this seems to be the wrong diagnosis of which of the competing hypothesis is better. Evelyn Adams won the lottery twice. Under the hypothesis that these events were causally unconnected and that each win was due to a random draw from the tickets purchased, the probability of the observations is very small. It is easy to construct hypotheses of Causal Connection that have much higher likelihoods. One of them says that her winning the first time was a random event, but that the occurrence of that first win guaranteed that she would win the next time. Another says that both lotteries were rigged so that she would win. This latter hypothesis has a likelihood than which none greater can be conceived; it has a likelihood of unity. The Law of Likelihood seems to endorse the naïve impulse to see conspiracies everywhere, to always think that a hypothesis of Causal Connection is better than the hypothesis of Mere Coincidence.

Bayesianism provides a natural solution to this type of problem for a wide range of cases. If prior probabilities can be defended by appeal to evidence, and aren't merely reflections of someone's subjective degrees of belief, then perhaps the likelihood advantage that conspiracy theories have can be overcome. Do we know that most state lotteries are fair? If so, this frequency data allows us to justify the assumption that the New Jersey lottery is probably fair. If the value of this defensible prior is high enough, we may be able to show that the conspiracy theory about Adams' double win has a low posterior probability even if it has a high likelihood.

30.6 The Limits of Bayesianism

The problem with this Bayesian solution is that there are lots of cases in which it isn't possible to back up assignments of prior probabilities with evidence and yet we still feel that there is something fishy about conspiracy theories and other hypotheses of causal connection.

In discussing the example of Wegener and continental drift, I noted that the hypothesis of Continental Drift has a much higher likelihood than the hypothesis of Continental Stasis: Pr(Data | Drift) >> Pr(Data | Stasis). However, this doesn't settle the matter of which hypothesis has the higher posterior probability. To decide that question, we must say something about the values of the prior probabilities, Pr(Drift) and Pr(Stasis). Geophysicists rejected Wegener's theory because they were sure that continents cannot plough through the ocean floor. Biologists and other friends of continental drift replied that this, or something like it, had to be possible, since the data are overwhelming. One aspect of the controversy that retarded the achievement of consensus was the way in which Wegener formulated his hypothesis. He could have restricted himself to the claim that the continents were once in contact, and not hazarded a guess about how they moved apart. He did not do this; as noted, he argued that the continents move across the ocean floor. He turned out to be right about the general claim, but wrong about the specifics. The continents don't move across the ocean floor. Rather, they and the ocean floor move together, the continents atop plates that slide across the viscous material that is deeper inside the earth.

A Bayesian will represent the disagreement between critics and defenders of continental drift by saying that they had different prior probabilities. Since the likelihoods overwhelmingly favor Drift over Stasis, the critics must have assigned to the drift hypothesis a prior probability that was incredibly small. Were they rational to do so? Or should they have assigned the hypothesis a somewhat larger prior, one that, though still small, allowed the data to give the drift hypothesis the higher posterior probability? It is hard to see how there can be an objective answer to that question. The prior probabilities were not estimated from frequency data. It's not as if a team of scientists visited a large number of planets, recording in each case whether the continents move, and then estimated from that data how probable it is that the continents move here on earth. Of course, there's another possible source of objective probabilities—ones that are derived from a well-confirmed theory. Did geophysicists have such a theory? If so, what probability did that theory entail for the hypothesis of continental drift? If the theory entails that continental drift is impossible, the Bayesian has a problem. The problem derives from the fact that a hypothesis assigned a prior probability of zero cannot have its probability increase, no matter what the evidence is. This is why Bayesians usually advise assigning priors of zero only to contradictions. Following this advice, we should decline to assign continental drift a prior of zero, even if our best confirmed theories say that drift is impossible. But what small prior should one then choose? If we choose a value that is extremely tiny, Drift will have a lower posterior probability than Stasis, even though Drift has the higher likelihood. If the prior probability is assigned a value that is a bit bigger, though still very small, Drift will end up with the larger posterior probability. No wonder the two communities were so divided. It is hard to see how the Bayesian can help decide what the correct assignment of prior probabilities is. Different groups of scientists had different degrees of belief; that appears to be all one can say.

Another scientific problem exhibits the same pattern. Consider the fact that the correlation of the phases of the moon and the tides were known for hundreds of years. It was not until Newton's theory of gravity that a systematic explanation of the correlation was developed. Newton's theory says that the two events are causally connected—the moon exerts a gravitational attraction on the earth's surface, with the result that there are tides. It is an objective matter that this hypothesis of causal connection has a higher likelihood than the hypothesis that says that it is a Mere Coincidence that the tides and the phases of the moon coincide: Pr(data | Newtonian Theory) >> Pr(data | Mere Coincidence). But does that mean that Newtonian theory is more probable than the hypothesis that the moon and the tides are causally unconnected? That depends on one's choice of priors. If Pr(Newtonian Theory) isn't enormously tiny, then Pr(Newtonian Theory | data) > Pr(Mere Coincidence | data). But if Newtonian theory is assigned a small enough prior, the theory will not be more probable than the hypothesis of Mere Coincidence. Unfortunately, there appears to be no objective basis for assigning priors in one way rather than the other.

Does a Bayesian analysis provide a convincing explanation of why Evelyn Adams' double win on the New Jersey lottery should be thought of as a Mere Coincidence? We need priors on the two hypotheses. Does any of us have frequency data on how often state lotteries, and the lottery in New Jersey specifically, are fixed? Surely if fixes occur, the parties will have every reason to prevent them from becoming public. How often they will succeed is another matter. My hunch is that the slogan "the truth will out" is too optimistic. In addition, how often the truth outs is more or less unknown. For this reason, we should be somewhat reluctant to interpret absence of evidence as evidence of absence in this instance.[7] I do not say that there is no objective basis for assigning prior probabilities here. Still, it would be nice if an analysis of this problem could be developed that did not require this. In other examples, the prospect for coming up with defensible priors for the candidate hypotheses is even more daunting.

The problem with Bayesianism isn't just about its use of priors. Its handling of likelihoods also raises questions when one or more of the hypotheses one wishes to consider is *composite*. It is perfectly clear what the probability of Adams' double win is, given the hypothesis that the two lotteries were fixed so that she would win. And it also is clear what the probability would be, given the hypothesis that each lottery was fair. But these two hypotheses are not exhaustive. Let us consider the complement of the first. What is the probability of Adams' double win if the lotteries were *not* fixed so as to ensure that she would win? There are many specific ways ($W_1, W_2, \ldots, W_n$) in which the lotteries could fail to be fixed to ensure Adams' double win. The likelihood of *Not-Fixed* is an average over all of these:

$$\Pr(\text{Adams wins both}|\text{Not-Fixed}) = \sum_i \Pr(\text{Adams wins both}|W_i)\Pr(W_i|\text{Not-Fixed})$$

It is the second product term in this summation that can be difficult to judge. Of course, if the goal is merely *psychological*—to describe how agents *actually* reason—this may not be an impediment. Perhaps people do have degrees of belief of the kind required. But if the goal is *normative*—to describe how we *ought* to reason—this can be a problem. In other examples, the prospect for coming up with defensible likelihoods for composite hypotheses is even more daunting.[8]

30.7 Models for a Larger Data Set

Imagine that we have data on all the people who bought tickets in all the New Jersey lotteries that have ever occurred, as well as information on who won what. Evelyn Adams' double win is part of this large data set, but only a small part. I want to consider a variety of models that might be offered for these multiple lotteries. What I mean by a "model" will be clarified in due course. To simplify discussion, I'll assume that there is just one winner in each lottery.

[7] There is an observation selection effect here; for discussion, see Sober (2004, 2009).

[8] See Griffiths and Tenenbaum (2007) for an interesting psychological study of how people actually think about coincidences that uses a Bayesian framework.

The first model I'll consider says that each lottery is fair—each ticket in a lottery has the same probability of winning:

(FAIR) If ticket t is purchased in lottery i $(1 \leq i \leq r)$, $\Pr(t \text{ wins}|t \text{ was purchased in lottery } i) = \alpha_i$.

The FAIR model is an r-fold conjunction:

$$\Pr(t \text{ wins}|t \text{ was purchased in lottery } 1) = \alpha_1.$$
$$\Pr(t \text{ wins}|t \text{ was purchased in lottery } 2) = \alpha_2.$$
$$\ldots$$
$$\Pr(t \text{ wins}|t \text{ was purchased in lottery } r) = \alpha_r.$$

By assigning a different parameter to each lottery, FAIR allows, *but does not require*, that the probability a ticket has of winning in one lottery differs from the probability a ticket has of winning in another. Notice also that this model does not say what the probability is of a ticket's winning any lottery. This model has r adjustable parameters, one for each lottery; each parameter "α_i" is bound to its own existential quantifier. The values of these probabilities must be estimated from the data. In each lottery i, there are n_i tickets sold and exactly one ticket was the winner. This means that the maximum likelihood estimate (the MLE) of α_i is $1/n_i$.

The second model I'll describe is more complicated than FAIR. It assigns a separate parameter to each player-lottery pair:

(PL) If ticket t is purchased in lottery i $(1 \leq i \leq r)$ by player j $(1 \leq j \leq s)$, $\Pr(t \text{ wins}|t \text{ was purchased in lottery } i \text{ by player } j) = \beta_{ij}$.

This model is a conjunction that contains rs conjuncts. It allows for the possibility that some or all the lotteries are unfair, but does not require this. The MLE of β_{ij} for player j on lottery i is 0 if the player lost, and $1/n_{ij}$ if the player won, where n_{ij} is the number of tickets the player purchased on that lottery.

The third model I'll consider is even more complicated. Like the one just described, it treats each player-lottery pair as a separate problem, but it introduces the possibility that different tickets purchased by the same player on the same lottery may have different probabilities of winning.

(PLT) If ticket t is the kth ticket purchased $(1 \leq k \leq n)$ in lottery i $(1 \leq i \leq r)$ by player j $(1 \leq j \leq s)$, $\Pr(t \text{ wins}|t \text{ is the } k\text{th ticket purchased in lottery } i \text{ by player } j) = \gamma_{ijk}$.

This model is a conjunction with rsn conjuncts. Notice that FAIR has the smallest number of parameters of the models described so far, and that PL and PLT both say that each lottery might be unfair but need not be.

The fourth and last model I'll consider (not that there aren't many others), involves circling back to the beginning to find a model that is even simpler than FAIR. FAIR allows that tickets in different lotteries may have different probabilities of winning. This is why that model has r parameters in it, one for each lottery. If we constrain tickets in all lotteries to have the same probability of winning, we obtain the following one-parameter model:

(ONE) If ticket t is purchased in any lottery, $\Pr(t \text{ wins} | t \text{ was purchased in a lottery}) = \delta$.

In a sense, this model says the lotteries have a greater degree of "fairness" than FAIR itself asserts. According to FAIR, players who buy a ticket in one lottery might have better odds than players who buy a ticket in another. The ONE model stipulates that this isn't so—every ticket in every lottery is in the same boat.

These different conceptualizations of how the lotteries work are "models" in the sense of that term that is standard in statistics. Each contains one or more adjustable parameters whose values can be estimated from the data. To clarify how these models are related to each other, let me describe two of their properties. First, notice that the models are nested; they are linked to each other by the relation of logical implication:

$$\text{ONE} \rightarrow \text{FAIR} \rightarrow \text{PL} \rightarrow \text{PLT}$$

Logically stronger models are special cases of models that are logically weaker. A stronger model can be obtained from a weaker one by stipulating that various parameters in the weaker model have equal values. Because of this, FAIR cannot be more probable than either PL or PLT, regardless of what the data are. Bayesians who want to argue that one of the simpler models has a higher prior or posterior probability than a model that is more complex might reply that the right way to set up models is to ensure that they are incompatible with each other; they should not be nested. This imperative requires that we compare ONE with FAIR*, PL*, and PLT*, where each of the starred models stipulates that different parameters must have different values. There is no logical barrier to stipulating that FAIR has a higher prior probability than either PL* or PLT*, but it is questionable whether there is a convincing reason to think that this stipulation is true. Is it really more probable that all tickets have exactly the same probability of winning a lottery than that they differ, if only by a little? I myself think it is very improbable that lotteries are *exactly* fair. Lotteries are like coins. I think that no coin is exactly fair. Coins in the real world have probabilities of landing heads that are *approximately* ½, not *exactly* ½. The other property of these models that I want to mention concerns the likelihoods they have when adjustable parameters are replaced by their maximum likelihood estimates. What I want to consider, for example, is not Pr(data | FAIR), but Pr[data | L(FAIR)], where L(FAIR) denotes the instance of FAIR obtained by assigning values to its parameters that make the data most probable. The point of interest here is that L(FAIR) can't have a higher likelihood than either L(PL) or L(PLT).[9] Increasing the number of adjustable parameters allows the resulting, more complex, model to fit the data better. In fact, the two most complex models, PL and PLT, are so complex that L(PL) and L(PLT) both say that Evelyn Adams was certain to win the two lotteries she did win, and that the winners of the other lotteries also had probabilities of unity of winning theirs. L(PLT) goes even farther; it says, not just that Adams was certain to win each of those two lotteries, but that it was a certainty that the tickets that won the two lotteries for her would do so. L(PL) doesn't go that far; if Adams purchased multiple tickets on one of the lotteries she won, L(PL) says that those tickets had equal probabilities of winning.

[9] L(FAIR) can't have a higher likelihood than L(PL*) or L(PLT*), either.

Comparing these models leads to a point that I think is of the first importance in our quest to understand how we should reason about coincidences. The naïve think that nothing is a Mere Coincidence. And the explanations they suggest for coinciding observations often seem to be very simple. When the naïve propose to explain Adams' double win by saying that the two lotteries were fixed, it would seem perverse to complain that this is a complicated explanation. What's so complicated about it? However, if we view this explanation as deriving from a model whose parameters are estimated from the data, and if we require that model to address a data set that is considerably more inclusive than these two facts about Adams, it turns out that the model that the naïve are implicitly using is vastly complex. They seem to be using a model that, when fitted to the data, says that each event that occurred had to occur. The hypothesis that all state lotteries have been FAIR is much simpler. Understanding the epistemic relevance of simplicity would throw light on the problem at hand.

30.8 Simplicity and Model Selection

Not only do we need to consider a larger data set instead of focusing exclusively on Adams' double win; we also must adjust our conception of what the goals are in model evaluation. The point is not to find a model that summarizes the data we have, but to find a model that will do a good job predicting data that we do not yet have. For example, suppose we were to use data on past New Jersey lotteries to compare models where our goal is to figure out which model will allow us to make the most accurate predictions about next year's lotteries. Of course, there's no getting around the Humean point that we have no assurance that future lotteries will play by the rules that governed past lotteries. But let us assume that this is true. How can we use the old data to estimate how well models will do in predicting new data?

Scientists who work on empirical problems by trying out multiple models inevitably learn that hugely complicated models often do a poor job predicting new data when fitted to old data. These models are able to *accommodate* the old data; as noted earlier, adding parameters to a model will allow it to fit the data better, and if M is sufficiently complex, Pr[old data | L(M)] = 1. However, Pr[new data | L(M)] will often be very low, or, more precisely, the distance between the predicted values and the observed values in the new data will often be great. This doesn't lead scientists to think that they should use the simplest possible model to make predictions.

Rather, some sort of trade-off is needed—the best model among the candidate models considered will embody the most nearly optimal trade-off between its fit to old data and its simplicity. How is that optimal balancing to be ascertained? Is it a matter of art, but not of science? Must young scientists simply work away at a given problem and gradually develop a feel for what works? Is this a reflection of the "tacit dimension" that Polanyi (1966) discussed? Well, there's no substitute for practical experience. However, there is, in addition, a body of results in mathematical statistics that shows that it is not a mere coincidence that very complicated models often make very inaccurate predictions. One central result in this literature is a theorem due to H. Akaike (1973), which says that

An unbiased estimate of the predictive accuracy of model M $\approx \log[\Pr(\text{data} \mid L(M))] - k$,

where k is the number of adjustable parameters in M. Akaike's theorem shows how good fit-to-data, as measured by the log-likelihood, improves expected predictive accuracy, while complexity, as measured by the number of adjustable parameters, diminishes that expectation. It also specifies a precise rate-of-exchange between log-likelihood and simplicity. It tells you how much of an improvement in fit-to-data is needed for the shift from a simpler to a more complex model to embody a net improvement in expected predictive accuracy (Forster and Sober 1994, Sober 2008).

Akaike's theorem is the basis for the Akaike Information Criterion (AIC), which scores a model by computing $-2[\log[\Pr(\text{data} \mid L(M))] - k]$; the best model will have the lowest AIC value. There are other model selection criteria on the market. Most of them are intended to help one identify models that are predictively accurate, and most of them include a penalty for complexity[10]; for discussion, see Burnham and Anderson (2002). There seems to be a broad consensus that different model selection criteria are appropriate for different inference problems.

If we use AIC to evaluate different models of the New Jersey lotteries, what will be the upshot? That will depend on the data. L(FAIR) will have a lower log-likelihood than L(LP) and L(LPT), but that doesn't ensure that FAIR is the worst of the three. The reason is that FAIR is far simpler than LP and LPT. It would not be surprising if FAIR scored better than these two more complicated models, but I cannot assert that this is true, since I have not looked at the data. However, the relevant epistemological point is visible without us having to carry out this set of calculations. FAIR may be a better model of the New Jersey lotteries than models like LP and LPT, which say that one or all of the lotteries may have been rigged, and this can be true even though L(FAIR) has a lower likelihood than L(LP) and L(LPT).

The model selection framework is not a magic bullet that will instantaneously convert the naïve into sophisticates. The naïve might reject the goal of predictive accuracy; they also may insist on focusing just on Adams' double win and refuse

[10] Cross validation makes no explicit mention of simplicity, but shares with AIC the goal of finding models that will be predictively accurate. It is interesting that there is a form of cross-validation ("take-one-out" cross validation) that is asymptotically equivalent with AIC (Stone 1977).

to consider the other data that constitute the history of the New Jersey Lottery. If they do so, they will have built a mighty fortress. If you look just at the double win, and don't want anything besides a hypothesis of maximum likelihood, there is no denying that the hypothesis that the two lotteries were both fixed to ensure that Adams would win beats the pants off the hypothesis that the two lotteries were fair.[11] But if you are prepared to ask the data to help you decide among the models just described, it may turn out that the FAIR model is superior to the PL and the PLT models. It is interesting that you don't need to evaluate the prior probabilities of PL and PLT to see what is wrong with these models. Indeed, since PL and PLT are consequences of FAIR, neither of these more complex models can have prior or posterior probabilities that are lower than the ones that attach to FAIR.

30.9 Conclusion

Having come this far—from probabilistic *modus tollens* to the Law of Likelihood to Bayesianism and then to model selection—let's return to an idea I mentioned towards the beginning. This is Diaconis and Mosteller's (1989, 859) *Law of Truly Large Numbers*, which says that "with a large enough sample, any outrageous thing is likely to happen." This principle implicitly assumes a certain type of model. As Diaconis and Mosteller are well aware, it isn't true in a suitably arranged deterministic model that any outrageous thing is likely to happen with enough trials, and the same point applies to many models that are probabilistic. The heuristic value of their principle is that it recommends that we look at the world in a certain way—we should use models that say that coinciding events can and do occur as Mere Coincidences, and have very high probabilities of doing so when the sample size is very large. But what are the rules of inference that recommend such models above others? The *Law of Truly Large Numbers* is not intended to address this question.

When two or more events are observed to coincide, the Law of Likelihood allows us to compare hypotheses of Mere Coincidence with hypotheses of Causal Connection, but often seems unable to identify a respect in which the first type of hypothesis is superior to the second. This is especially clear when the Causal Connection Hypothesis is deterministic and the Mere Coincidence hypothesis is probabilistic. The Bayesian response to this problem is to assign prior probabilities. Sometimes these can be justified by appeal to evidence; at other times, they seem to be merely subjective. It is in the latter kind of case that model selection criteria seem like a breath of fresh air.

[11] It might be suggested that the hypothesis that the two lotteries were fixed to ensure that Adams would win is a hypothesis that would occur to you only *after* you observe Adams' double win, and that it is a rule of scientific inference that hypotheses must be formulated *before* the data are gathered to test them. This temporal requirement is often invoked in frequentist statistics. For discussion, see Hitchcock and Sober (2004). It is a point in favor of the model selection approach that one does not have to invoke this temporal requirement to explain what is wrong with the PL and the PLT models.

Some years ago, cognitive psychologists discussed the phenomenon of "hot hands" in sports. Everyone with even the most superficial familiarity with professional basketball believes that players occasionally have "hot hands." When players are hot, their chance of scoring improves, and team-mates try to feed the ball to them. However, when Gilovich et al. (1985) did a statistical analysis of scoring patterns in the NBA, they concluded that one cannot reject the null hypothesis that each player has a constant probability of scoring throughout the season. These and many other statistically sophisticated scientists concluded that belief in hot hands is a "cognitive illusion." A scoring streak is not due to the player's getting hot, but is a Mere Coincidence.[12] Basketball mavens reacted to this statistical pronouncement with total incredulity.

What would a Bayesian analysis of this problem look like? Surely we have lots of evidence that physical injury, influenza, upset stomach, lack of sleep, and migraine impair athletic performance. The idea that a player's probability of scoring through the season is *absolutely constant* should therefore be assigned a very low prior probability. For this reason, Bayesianism seems predestined to side with common sense on this issue. I do not see this as a defect in Bayesianism, nor do I have any sympathy with the argument that defends the null hypothesis by pointing out that the data do not sanction its rejection. Is this another case of probabilistic *modus tollens'* rearing its ugly head? In any event, the model selection framework provides a very different and useful perspective.

Recall that the goal in model selection is to find models that will be predictively accurate. It is an important philosophical fact about this framework that we can have *evidence* that a model *known* to be false will be a better predictor than a model *known* to be true (Sober 2002, 2008, Forster and Sober 2011). Bayesians are right to say that the null hypothesis has very low prior and posterior probabilities. The idea that players never waiver in their scoring probabilities, even a little, *is* preposterous. However, this doesn't settle which model will make the most accurate predictions. Presumably, the truth about basketball players is very complex. Their scoring probabilities change as subtle responses to a large number of interacting causes. Given this complexity, players and coaches may make better predictions by relying on simplified models. Hot hands may be a reality, but trying to predict when players have hot hands may be a fool's errand.

Acknowledgements I thank Matthew Kopec, Ellery Eells, Arthur Fine, Malcolm Forster, George Gale, Michael Goldsby, Daniel Hausman, Stephen Leeds, Wouter Meijs, David Myers, Joshua Tenenbaum, and Naftali Weinberger for helpful discussion, and Nick Harding for permitting me to reprint his cartoon.

[12] See Wardrop (1999) for a skeptical assessment of Gilovich et al.'s analysis. Wardrop argues that Gilovich et al. tested hypotheses about *correlation* (whether a player's probability of scoring on a given shot if he scored on earlier shots is greater than his probability of scoring if he missed previously), but did not assess the issue of *stationarity* (maybe a player's probability of scoring suddenly shifts from one value to another). Wardrop suggests that the latter may be the relevant consideration.

References

Akaike, Hirotogu. 1973. Information theory as an extension of the maximum likelihood principle. In *Second international symposium on information theory*, eds. Boris Petrov and Frigyes Csáki, 267–281. Budapest: Akademiai Kiado.

Burnham, Kenneth, and David Anderson. 2002. *Model selection and inference – A practical information-theoretic approach*, 2nd ed. New York: Springer.

Cartwright, Nancy. 1994. *Nature's capacities and their measurement.* Oxford: Oxford University Press.

Darwin, Charles. 1859. *On the origin of species.* London: John Murray.

Darwin, Charles. 1871. *The descent of man and selection in relation to sex.* London: Murray.

Diaconis, Persi, and Frederick Mosteller. 1989. Methods of studying coincidences. *Journal of the American Statistical Association* 84: 853–861.

Edwards, Anthony. 1972. *Likelihood.* Cambridge: Cambridge University Press.

Fisher, Ronald Aylmer. 1959. *Statistical methods and scientific inference*, 2nd ed. New York: Hafner.

Fitelson, Branden. 1999. The plurality of bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science* 66: S362–S378.

Forster, Malcolm, and Elliott Sober. 1994. How to tell when simpler, more unified, or less *Ad Hoc* theories will provide more accurate predictions. *British Journal for the Philosophy of Science* 45: 1–36.

Forster, Malcolm, and Elliott Sober. 2011. AIC scores as evidence – A bayesian interpretation. In *the philosophy of statistics*, eds. Malcolm Forster and Prasanta Bandyopadhyay, 535–549. Dordrecht: Kluwer.

Gilovich, Thomas, Robert Valone, and Amos Tversky. 1985. The hot hand in basketball – On the misperception of random sequences. *Cognitive Psychology* 17: 295–314.

Griffiths, Thomas, and Joshua Tenenbaum. 2007. From mere coincidences to meaningful discoveries. *Cognition* 103: 180–226.

Hacking, Ian. 1965. *The logic of statistical inference.* Cambridge: Cambridge University Press.

Hitchcock, Christopher, and Elliott Sober. 2004. Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science* 55: 1–34.

Littlewood, John. 1953. *A mathematician's miscellany.* London: Methuen.

Myers, David G. 2002. *Intuition – Its powers and perils.* New Haven, CT: Yale University Press.

National Commission on the Future of DNA Evidence. 2000. *The future of forensic DNA testing: Predictions of the Research and Development Working Group.* U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Pearl, Judea. 2000. *Causality – Models, reasoning, inference.* New York: Cambridge University Press.

Polanyi, Michael. 1966. *The tacit dimension.* New York: Doubleday.

Reichenbach, Hans. 1956. *The direction of time.* Berkeley, CA: University of California Press.

Royall, Richard. 1997. *Statistical evidence – A LIKELIHOOD Paradigm.* London: Chapman and Hall.

Sober, Elliott. 1988a. The principle of the common cause. In *Probability and causation: Essays in honor of Wesley Salmon*, ed. James Fetzer, 211–228. Dordrecht: Reidel.

Sober, Elliott. 1988b. *Reconstructing the past – Parsimony, evolution, and inference.* Cambridge, MA: MIT Press.

Sober, Elliott. 2001. Venetian sea levels, British Bread Prices, and the principle of the common cause. *British Journal for the Philosophy of Science* 52: 1–16.

Sober, Elliott. 2002. Instrumentalism, parsimony, and the Akaike framework. *Philosophy of Science* 69: S112–S123.

Sober, Elliott. 2004. The design argument. In *Blackwell guide to the philosophy of religion*, ed. William Mann, 117–147. Oxford: Blackwell.

Sober, Elliott. 2008. *Evidence and evolution – The logic behind the science*. Cambridge: Cambridge University Press.
Sober, Elliott. 2009. Absence of evidence and evidence of absence – Evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads. *Philosophical Studies* 143: 63–90.
Sober, Elliott. 2011. *Did Darwin write the origin backwards?* Amherst, NY: Prometheus Books.
Spirtes, Peter, Clark Glymour, and Richard Scheines. 2001. *Causality, prediction, and search*. Cambridge, MA: MIT Press.
Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B* 39: 44–47.
Wardrop, Robert. 1999. *Statistical tests for the hot-hand in basketball in a controlled setting*. Technical Report, Department of Statistics, University of Wisconsin, Madison. http://hot-hand.behaviouralfinance.net/Ward99.pdf. Accessed 2 Apr 2007.
Wegener, Alfred. 1924. *The origin of continents and oceans: Translation of German*, 3rd ed. London: Methuen.
Woodward, James. 2003. *Making things happen*. Oxford: Oxford University Press.

Chapter 31
Stopping Rules and Data Monitoring in Clinical Trials

Roger Stanev

31.1 Introduction

Stopping rules—rules dictating when to stop accumulating data and start analyzing it for the purposes inferring from the experiment—divide Bayesians, Likelihoodists and classical statistical approaches to inference. Although the relationship between Bayesian philosophy of science and stopping rules can be complex (cf. Steel 2003), in general, Bayesians regard stopping rules as irrelevant to what inference should be drawn from the data. This position clashes with classical statistical accounts. For orthodox statistics, stopping rules do matter to what inference should be drawn from the data. "The dispute over stopping rule is far from being a marginal quibble, but is instead a striking illustration of the divergence of fundamental aims and standards separating Bayesians and advocates of orthodox statistical methods" (Steel 2004, 195).

But philosophers who subscribe, on theoretical grounds, to particular principles of statistical inference need to recognize the limitations of the statistical approach they endorse when it comes to important matters, such as the conduct of randomized clinical trials (RCTs). In broadest terms, I am concerned with the following problem: what if no single statistical approach is best-suited to address all the necessary demands of clinical research? The paper focus on a specific version of this problem: the apparent inability of existing statistical approaches to accommodate two such demands. The first is that RCTs incorporate some basic stopping rule, and the second is that RCTs incorporate policies for early termination (at times in violation of the basic stopping rule). While many statistical approaches can meet one of these demands, no extant approach appears capable of meeting both. I suggest that this type of predicament requires new ways of thinking about the problem in order to give credit to distinct approaches where it might be due. Rather than solving the problem by formulating yet another universal paradigm for statistical inference,

R. Stanev (✉)
Department of Philosophy, University of British Columbia, Vancouver, BC, Canada
e-mail: rstanev@interchange.ubc.ca