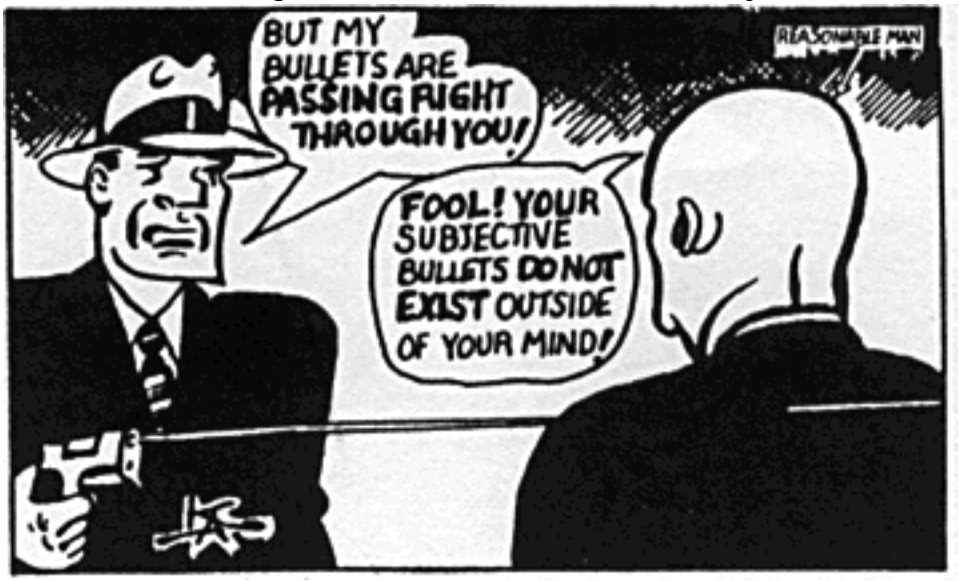
Subjective Probability



(The Real Thing)

Richard Jeffrey

©2001

December 4, 2001

Contents

1 Probability Primer (Statics)
1.1 Bets and Probabilities
1.2 Why Probabilities are Additive
1.3 Laws of Probability
1.4 Conditional Probability
1.5 Why '|' Cannot be a Connective
1.6 Bayes's Theorem
1.7 Independence
1.8 Exchangeability
1.9 Problems

2 Probability Dynamics; Learning
2.1 Conditioning
2.2 Probabilities from Statistics
2.3 Generalized Conditioning
2.4 "Real" Probability

3 Testing Scientific Theories
3.1 Huygens on Methodology
3.2 Quantifications of Confirmation
3.3 Observation and Sufficiency
3.4 Leverrier on Neptune
3.5 Multiple Uncertainties
3.6 Dorling on the Duhem Problem
3.6.1 Einstein/Newton, 1919
3.6.2 Bell's Inequalities: Holt/Clauser
3.6.3 Laplace/Adams
3.6.4 Dorling's Conclusions
3.7 Old News Explaned
3.8 Problems

4 Expectation Primer
4.1 Probability and Expectation
4.2 Conditional Expectation
4.3 Laws of Expectation
4.4 Physical Analogies; Mean and Median
4.5 Variance; Chebyshev-Markov Inequalities
4.6 Two Easy Laws of Large Numbers

Preface

The theory of probability was a mid-17th century artifact originating in a famous correspondence between Fermat and Pascal—a correspondence on which Huygens based a widely read textbook: *On Calculating in Games of Luck* (1657). Although probabilistic judgment may have had a long prehistory in a widespread human practice of betting at odds, the theoretical framework did not exist until those people cobbled it together. It remains in use today, much as in Huygens's book.

The version of Bayesianism presented here ("Radical Probabilism") sees making up your mind within that framework as a matter of designing or borrowing probability assignments to hypotheses about the real world, or, anyway, features of such assignments—as when you/we adopt the feature of equiprobability of two hypotheses without yet adopting any particular value for that probability. In large part this is a matter of artifice, a family of arts of probabilistic judgment. And that includes making up our minds about how to change our minds.

With the founders (so I think), and certainly with Frank Ramsey and Bruno de Finetti, who in the 20th century revived the view of probability as a mode of judgment, I see the probability calculus as a logic. Mere logic does not tell us how to make up our minds—how to design, or where to borrow, the needed features. It does help us spot inconsistencies within projected mental makeups, projected systems of features, but fails to underwrite the further gradations we make in classifying survivors of the first cut as reasonable or not, or in grading them as more or less reasonable. These finer distinctions separate cases on the basis of standards rooted in our actual appreciation of ongoing methodological experience.

The basic technical ideas, floated in chapters 1 and 4, are applied elsewhere to troublesome questions about scientific method and to problems of collaborative judgment of the sort found in medical diagnosis and decision-making.

Do send me corrections and other suggestions.

— Richard Jeffrey
(dickjeff@princeton.edu)

Chapter 1

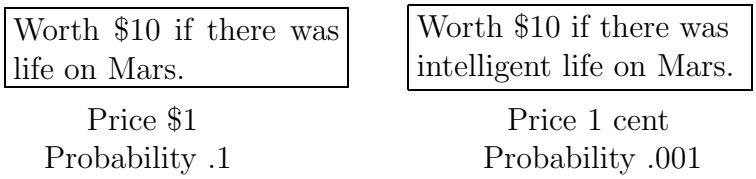
Probability Primer (Statics)

Yes or no: was there once life on Mars? We can't say. What about intelligent life? That seems most unlikely, but again, we can't really say. The simple yes-or-no framework has no place for shadings of doubt, no room to say that we see intelligent life on Mars as far less probable than life of a possibly very simple sort. Nor does it let us express exact probability judgments, if we have them. We can do better.

1.1 Bets and Probabilities

What if you were able to say exactly what odds you would give on there having been life, or intelligent life, on Mars? That would be a more nuanced form of judgment, and perhaps a more useful one. Suppose your odds were 1:9 for life, and 1:999 for intelligent life, corresponding to probabilities of 1/10 and 1/1000, respectively. (Odds m:n correspond to probability $\frac{m}{m+n}$.) That means you would see no special advantage for either player in risking one dollar to gain nine in case there was once life on Mars; and it means you would see an advantage on one side or the other if those odds were shortened or lengthened. And similarly for intelligent life on Mars when the risk is a thousandth of the same ten dollars (1 cent) and the gain is 999 thousandths ($9.99).

Here is another way of saying the same thing: You would think a price of one dollar just right for a ticket worth ten if there was life on Mars and nothing if there was not, but you would think a price of only one cent right if there would have had to have been intelligent life on Mars for the ticket to be worth ten dollars. These are the two tickets:



So if you have an exact judgmental probability for truth of a hypothesis, it corresponds to your idea of the dollar value of a ticket that is worth 1 unit or nothing, depending on whether the hypothesis is true or false. (For the hypothesis of mere life on Mars the unit was $10; the price was a tenth of that.)

Of course you need not have exact judgmental probability regarding life on Mars, or intelligent life there. Still, we know that any probabilities anyone might think acceptable for those two hypotheses ought to satisfy certain rules, e.g., that the first cannot be less than the second. That is because the second hypothesis implies the first. (See the implication rule at the end of sec. 1.3 below.) In sec. 1.2 we turn to the question of what the laws of judgmental probability are, and why. Meanwhile, take some time with the following questions, as a way of getting in touch with some of your own ideas about probability. Afterward, read the discussion that follows.

Questions

1 A vigorously flipped thumbtack will land on the sidewalk. Is it reasonable for you to have a probability for the hypothesis that it will land point up?

2 An ordinary coin is to be tossed twice in the usual way. What is your probability for the head turning up both times?
(a) 1/3, because 2 heads is one of three possibilities: 2, 1, 0 heads?
(b) 1/4, because 2 heads is one of four possibilities: HH, HT, TH, TT?

3 There are three coins in a bag: ordinary, two-headed, and two-tailed. One is shaken out onto the table and lies head up. What should be your probability that it's the two-headed one:
(a) 1/2, since it can only be two-headed or normal?
(b) 2/3, because the other side could be the tail of the normal coin, or either side of the two-headed one? (Suppose the sides have microscopic labels.)

4 *It's a goy!*[1]
(a) As you know, about 49% of recorded human births have been girls. What

[1]This is an adaptation of Nelson Goodman's "grue" paradox, in his *Fact, Fiction and Forecast,* 4th ed., Harvard U. P., 1983, pp. 73-4.

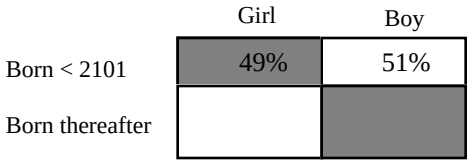
is your judgmental probability that the first child born after time t (say, $t =$ the beginning of the 21st century, GMT) will be a girl?
(b) A *goy* is defined as a girl born before t or a boy born thereafter. As you know, about 49% of recorded human births have been goys. What is your judgmental probability that the first child born in the 22nd century will be a goy?

Discussion

1 Surely it is reasonable to suspect that the geometry of the tack gives one of the outcomes a better chance of happening than the other; but if you have no clue about which of the two has the better chance, it may well be reasonable to have judgmental probability 1/2 for each. Evidence about the chances might be given by statistics on tosses of similar tacks, e.g., if you learned that in 20 tosses there were 6 up's you might take the chance of up to be in the neighborhood of 30%; and whether or not you do that, you might well adopt 30% as your judgmental probability for up on the next toss.

2,3 These questions are meant to undermine the impression that judgmental probabilities can be based on analysis into cases in a way that does not already involve probabilistic judgment (e.g., the judgment that the cases are equiprobable). In either problem you can arrive at a judgmental probability by trying the experiment (or a similar one) often enough, and seeing the statistics settle down close enough to 1/2 or to 1/3 to persuade you that more trials will not reverse the indications. In each these problems it is the finer of the two suggested analyses that happens to make more sense; but any analysis can be refined in significantly different ways, and there is no point at which the process of refinement has to stop. (Head or tail can be refined to head–facing–north or head–not–facing–north or tail.) Indeed some of these analyses seem more natural or relevant than others, but that reflects the probability judgments we bring with us to the analyses.

4 *Goys and birls*



Shaded: Goy. Blank: Birl.

This question is meant to undermine the impression that judgmental probabilities can be based on frequencies in a way that does not already involve

judgmental probabilities. Since all girls born so far have been goys, the current statistics for girls apply to goys as well: these days, about 49% of human births are goys. Then if you read probabilities off statistics in a straightforward way your probability will be 49% for each of these hypothesis:

(1) The first child born after t will be a girl.

(2) The first child born after t will be a goy.

Thus pr(1)+pr(2)=98%. But it is clear that those probabilities should sum to 100%, since (2) is logically equivalent to

(3) The first child born after t will be a boy,

and pr(1)+pr(3) = 100%. To avoid this contradiction you must decide which statistics are relevant to $pr(1)$: the 49% of girls born before 2001, or the 51% of boys. And that is not a matter of statistics but of judgment—no less so because we would all make the same judgment: the 51% of boys.

1.2 Why Probabilities are Additive

Authentic tickets of the Mars sort are hard to come by. Is the first of them really worth $10 to me if there was life on Mars? Probably not. If the truth is not known in my lifetime, I cannot cash the ticket even if it is really a winner. But some probabilities are plausibly represented by prices, e.g., probabilities of the hypotheses about athletic contests and lotteries that people commonly bet on. And it is plausible to think that the general laws of probability ought to be the same for all hypotheses—about planets no less than about ball games. If that is so, we can justify laws of probability if we can prove all betting policies that violate them to be inconsistent. Such justifications are called "Dutch book arguments".[2] We shall give a Dutch book argument for the requirement that probabilities be additive in this sense:

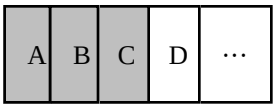
> **Finite Additivity.** The probability of a hypothesis that can be true in a finite number of incompatible ways is the sum of the probabilities of those ways.

[2]In British racing jargon a *book* is the set of bets a bookmaker has accepted, and a book *against* someone—a "Dutch" book—is one the bookmaker will suffer a net loss on no matter how the race turns out. I follow Brian Skyrms in seeing F. P. Ramsey as holding Dutch book arguments to demonstrate actual inconsistency. See Ramsey's 'Truth and Probability' in his *Philosophical Papers,* D. H. Mellor, ed.: Cambridge, 1990.

EXAMPLE 1, **Finite additivity.** The probability p of the hypothesis

(H) A sage will be elected

is $q+r+s$ if exactly three of the candidates are sages and their probabilities of winning are q, r, and s. In the following diagram, $A, B, C, D, E, \ldots$ are the hypotheses that the various different candidates win—the first three being the sages. The logical situation is diagrammed as follows, where the points in the big rectangle represent all the ways the election might come out, specified in minute detail, and the small rectangles represent the ways in which the winner might prove to be A, or B, or C, or D, etc.



H true H false

Probabilities of cases A, B, C, D,...
are q, r, s, t,..., respectively.

Dutch Book Argument for Finite Additivity.

For definiteness we suppose that the hypothesis in question is true in three cases, as in example 1; the argument differs inessentially for other examples, with other finite numbers of cases. Now consider the following four tickets.

Worth \$1 if H is true.	Price \$$p$
Worth \$1 if A is true.	Price \$$q$
Worth \$1 if B is true.	Price \$$r$
Worth \$1 if C is true.	Price \$$s$

Suppose I am willing to buy or sell any or all of these tickets at the stated prices. Why should p be the sum $q+r+s$? Because no matter what it is worth —\$1 or \$0—the ticket on H is worth exactly as much as the tickets on A, B, C together. (If H loses it is because A, B, C all lose; if H wins it is because exactly one of A, B, C wins.) Then if the price of the H ticket is different from the sum of the prices of the other three, I am inconsistently placing different values on one and the same contract, depending on how it is presented.

If I am inconsistent in that way, I can be fleeced by anyone who will ask me to buy the H ticket and sell or buy the other three depending on whether p is more or less than $q+r+s$. Thus, no matter whether the equation $p = q+r+s$ fails because the left–hand side is more or less than the right, a book can be made against me. That is the Dutch book argument for additivity when the number of ultimate cases under consideration is finite. The talk about being fleeced is just a way of dramatizing the inconsistency of any policy in which the dollar value of the ticket on H is anything but the sum of the values of the other three tickets: to place a different value on the three tickets on A, B, C from the value you place on the H ticket is to place different values on the same commodity bundle under two demonstrably equivalent descriptions.

When the number of cases is infinite, a Dutch book argument for additivity can still be given—provided the infinite number is not too big! It turns out that not all infinite sets are the same size:

EXAMPLE 2, **Cantor's Diagonal Argument.** The sets of positive integers (I^+'s) cannot be counted off as first, second, ..., with each such set appearing as n'th in the list for some finite positive integer n. This was proved by Georg Cantor (1895) as follows. Any set of I^+'s can be represented by an enless string of plusses and minuses ("signs"), e.g., the set of even I^+'s by the string $-+-+\ldots$ in which plusses appear at the even numbered positions and minuses at the odd, the set $\{2, 3, 5, 7, \ldots\}$of prime numbers by an endless string that begins $-++-+-+$, the set of all the I^+'s by an endless string of plusses, and the set of no I^+'s by an endless string of minuses. Cantor proved that no list of endless strings of signs can be complete. He used an amazingly simple method ("diagonalization") which, applied to any such list, yields an endless string $\bar{d}$ of signs which is not in that list. Here's how. For definiteness, suppose the first four strings in the list are the examples already given, so that the list has the general shape

$$\begin{array}{c} s_1 : - + - + \ldots \\ \mathrm{s}_2 : - + + - \ldots \\ \mathrm{s}_3 : + + + + \ldots \\ \mathrm{s}_4 : - - - - \ldots \\ \text{etc.} \end{array}$$

Define the *diagonal* of that list as the string d consisting of the first sign in s_1, the second sign in s_2, and, in general, the n'th sign in s_n :

$$- + + - \ldots$$

And define the *antidiagonal* $\bar{d}$ of that list as the result $\bar{d}$ of reversing all the signs in the diagonal,

$$\bar{d} : + - - + \ldots$$

In general, for any list $s_1, s_2, s_3, s_4 \ldots$, $\bar{d}$ cannot be any member s_n of the list, for, by definition, the n'th sign in $\bar{d}$ is different from the n'th sign of s_n, whereas if $\bar{d}$ were some s_n, those two strings would have to agree, sign by sign. Then the set of I^+'s defined by the antidiagonal of a list cannot be in that list, and therefore no list of sets of I^+'s can be complete.

> **Countability.** A countable set is defined as one whose members (if any) can be arranged in a single list, in which each member appears as the n'th item for some finite n.

Of course any finite set is countable in this sense, and some infinite sets are countable.

An obvious example of a countably infinite set is the set $I^+ = \{1, 2, 3, \ldots\}$ of all positive whole numbers.

A less obvious example is the set I of all the whole numbers, positive, negative, or zero: $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$. The members of this set can be rearranged in a single list of the sort required in the definition of countability: $0, 1, -1, 2, -2, 3, -3, \ldots$. So the set of all the whole numbers is countable. Order does not matter, as long as every member of I shows up somewhere in the list.

Example 3, **Countable additivity.** In example 1, suppose there were an endless list of candidates, including no end of sages. If H says that a sage wins, and $A_1, A_2, \ldots$ identify the winner as the first, second, ... sage, then an extension of the law of finite additivity to countably infinite sets would be this:

> **Countable Additivity.** The probability of a hypothesis H that can be true in a countable number of incompatible ways $A_1, A_2 \ldots$ is the sum $pr(H) = pr(A_1) + pr(A_2) + \ldots$ of the probabilities of those ways.

This equation would be satisfied if the probability of one or another sage's winning were $pr(H) = 1/2$, and the probabilities of the first, second, third, etc. sage's winning were 1/4, 1/8, 1/16, etc., decreasing by half each time.

1.2.1 Dutch book argument for countable additivity.

Consider the following infinite array of tickets, where the mutually incompatible A's collectively exhaust the ways in which H can be true (as in example 3).[3]

Pay the bearer \$1 if H is true	Price \$$pr(H)$
Pay the bearer \$1 if A_1 is true	Price \$$pr(A_1)$
Pay the bearer \$1 if A_2 is true	Price \$$pr(A_2)$
...	...

Why should my price for the first ticket be the sum of my prices for the others? Because no matter what it is worth —\$1 or \$0—the first ticket is worth exactly as much as all the others together. (If H loses it is because the others all lose; if H wins it is because exactly one of the others wins.) Then if the first price is different from the sum of the others, I am inconsistently placing different values on one and the same contract, depending on how it is presented.

Failure of additivity in these cases implies inconsistency of valuations: a judgment that certain transactions are at once (1) reasonable and (2) sure to result in an overall loss. Consistency requires additivity to hold for countable sets of alternatives, finite or infinite.

1.3 Probability Logic

The simplest laws of probability are the consequences of finite additivity under this additional assumption:

> Probabilities are real numbers in the range from 0 to 1, with the endpoints reserved for certainty of falsehood and of truth, respectively.

[3]No matter that there is not enough paper in the universe for an infinity of tickets. One small ticket can save the rain forest by doing the work of all the A tickets together. This eco-ticket will say: 'For each positive whole number n, pay the bearer \$1 if A_n is true

This makes it possible to read probability laws off diagrams, much as we read ordinary logical laws off them. Let's see how that works for the ordinary ones, beginning with two surprising examples (where "iff" means *if and only if*):

De Morgan's Laws:

(1) $\neg(G \wedge H) = \neg G \vee \neg H$ ("Not both true iff at least one false")
(2) $\neg(G \vee H) = \neg G \wedge \neg H$ ("Not even one true iff both false")

Here the bar, the wedge and juxtaposition stand for *not, or* and *and.* Thus, if G and H are two hypotheses,

$G \wedge H$ (or GH) says that they are both true: G AND H

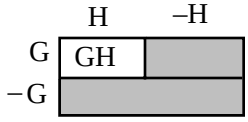
$G \vee H$ says that at least one is true: G OR H

$\neg G$ (or $-G$ or $\bar{G}$) says that G is false: NOT G

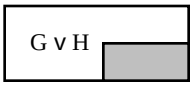
In the following diagrams for De Morgan's laws the upper and lower rows represent G and $\neg G$ and the left and right-hand columns represent H and $\neg H$. Now if R and S are any regions, $R \wedge S$ (or 'RS'), is their intersection, $R \vee S$ is their union, and $\neg R$ is the whole big rectangle except for R.

Diagrams for De Morgan's laws (1) and (2):

(1) Shaded: $\neg(G \wedge H) = \neg G \vee \neg H$



(2) Shaded: $\neg(G \vee H) = \neg G \wedge \neg H$



Adapting such geometrical representations to probabilistic reasoning is just a matter of thinking of the probability of a hypothesis as its region's area, assuming that the whole rectangle, $H \vee \neg H$ $(= G \vee \neg G)$, has area 1. Of course the empty region, $H \wedge \neg H$ $(= G \wedge \neg G)$, has area 0. It is useful to denote those two extreme regions in ways independent of any particular hypotheses H, G. Let's call them $\top$ and $\bot$:

Logical Truth. $\top = H \vee \neg H = G \vee \neg G$
Logical Falsehood. $\bot = H \wedge \neg H = G \wedge \neg G$

We can now verify some further probability laws informally, in terms of areas of diagrams.

$$\textbf{Not}: \ pr(\neg H) = 1 - pr(H)$$

Verification. The non-overlapping regions H and $\neg H$ exhaust the whole rectangle, which has area 1. Then $pr(H) + pr(\neg H) = 1$, so $pr(\neg H) = 1 - pr(H)$.

$$\textbf{Or:} \ pr(G \vee H) = pr(G) + pr(H) - pr(G \wedge H)$$

Verification. The $G \vee H$ area is the G area plus the H area, except that when you simply add $pr(G) + pr(H)$ you count the $G \wedge H$ part twice. So subtract it on the right-hand side.

The word 'but'—a synonym for 'and'—may be used when the conjunction may be seen as a contrast, as in 'it's green but not healthy', $G \wedge \neg H$:

$$\textbf{But Not:} \ pr(G \wedge \neg H) = pr(G) - pr(G \wedge H)$$

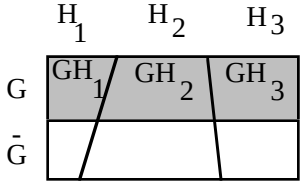
Verification. The $G \wedge \bar{H}$ region is what remains of the G region after the $G \wedge H$ part is deleted.

$$\textbf{Dyadic Analysis:} \ pr(G) = pr(G \wedge H) + pr(G \wedge \neg H)$$

Verification. See the diagram for De Morgan (1). The G region is the union of the nonoverlapping $G \wedge H$ and $G \wedge \neg H$ regions.

In general, there is a rule of n-adic analysis for each n, e.g., for n=3:

Triadic Analysis: If H_1, H_2, H_3 *partition* $\top$,[4] then
$pr(G) = pr(G \wedge H_1) + pr(G \wedge H_2) + pr(G \wedge H_3)$.

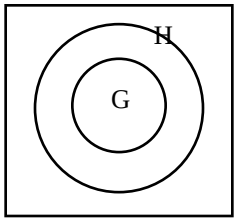


The next rule follows immediately from the fact that logically equivalent hypotheses are always represented by the same region of the diagram—in view of which we use the sign '=' of identity to indicate logical equivalence.

[4]This means that the H's are mutually exclusive ($H_1 \wedge H_2 = H_1 \wedge H_3 = H_2 \wedge H_3 = \bot$) and collectively exhaustive ($H_1 \vee H_2 \vee H_3 = \top$).

Equivalence: If $H = G$, then $pr(H) = pr(G)$.
(Logically equivalent hypotheses are equiprobable.)

Finally: To be implied by G, the hypothesis H must be true in every case in which G is true. Diagramatically, this means that the G region is entirely included in the H region. Then if G implies H, the G region can have no larger an area than the H region.



Implication: If G implies H, then $pr(G) \leq pr(H)$.

1.4 Conditional Probability

We identified your ordinary (unconditional) probability for H as the price representing your valuation of the following ticket

Worth \$1 if H is true.	Price: $\$pr(H)$

Now we identify your conditional probability for H given D as the price representing your valuation of this ticket:

(1)	Worth \$1 if $D \wedge H$ is true, worth $\$pr(H\|D)$ if D is false.	Price: $\$pr(H\|D)$

The old ticket represented a simple bet on H; the new one represents a conditional bet on H—a bet that is called off (the price of the ticket is refunded) in case the condition D fails. If D and H are both true, the bet is on and you win, the ticket is worth \$1. If D is true but H is false, the bet is on and you lose, the ticket is worthless. And if D is false, the bet is off, you get your $\$pr(H|D)$ back as a refund.

With that understanding we can construct a Dutch book argument for the following rule, which connects conditional and unconditional probabilities:

Product Rule: $pr(H \wedge D) = pr(H|D)pr(D)$

Dutch Book Argument for the Product Rule.[5] Imagine that you own three tickets, which you can sell at prices representing your valuations. The first is ticket (1) above. The second and third are the following two, which represent unconditional bets of \$1 *on* HD and of $\$pr(H|D)$ *against* D,

(2) $\boxed{\text{Worth \$1 if } H \wedge D \text{ is true.}}$ Price: $\$pr(H \wedge D)$

(3) $\boxed{\text{Worth } pr(H|D) \text{ if } D \text{ is false.}}$ Price: $\$pr(H|D)pr(\bar{D})$

Bet (3) has a peculiar payoff: not a whole dollar, but only $\$pr(H|D)$. That is why its price is not the full $\$pr(\neg D)$ but only the fraction $pr(\neg D)$ of the $\$pr(H|D)$ that you stand to win. This payoff was chosen to equal the price of the first ticket, so that the three fit together into a neat book:

Observe that in every possible case regarding truth and falsity of H and D the tickets (2) and (3) together have the same dollar value as ticket (1). (You can verify that claim with pencil and paper.) Then there is nothing to choose between ticket (1) and tickets (2) and (3) together, and therefore it would be inconsistent to place different values on them. Thus, your price for (1) ought to equal the sum of your prices for (2) and (3):

$$pr(H|D) = pr(H \wedge D) + pr(\neg D)pr(H|D)$$

Now set $pr(\neg D) = 1 - pr(D)$, multiply through, cancel $pr(H|D)$ from both sides and solve for $pr(H \wedge D)$. The result is the product rule. To violate that rule is to place different values on the same commodity bundle in different guises: (1), or the package (2, 3).

The product rule is more familiar in a form where it is solved for the conditional probability $pr(H|G)$:

Quotient Rule: $pr(H|D) = \dfrac{pr(H \wedge D)}{pr(D)}$, provided $pr(D) > 0$.

Graphically, the quotient rule expresses $pr(H|D)$ as the fraction of the D region that lies inside the H region. It is as if calculating $pr(H|D)$ were a

[5] Adapted from Bruno de Finetti, 'Foresight: Its Logical Laws and Subjective Sources', reprinted in *Studies in Subjective Probability,* 2nd ed., edited by Henry Kyburg, Jr. and Howard Smokler (Huntington, N.Y.: Robert E. Krieger, 1980).

matter of trimming the whole $D \vee \neg D$ rectangle down to the D part, and using that as the new unit of area.

The quotient rule is often called the definition of conditional probability. It is not. If it were, we could never be in the position we are often in, of making a conditional judgment—say, about a coin that may or may not be tossed will land—without attributing some particular positive value to the condition: $pr(\text{head} \mid \text{tossed}) = 1/2$ even though

$$\frac{pr(\text{head} \wedge \text{tossed})}{pr(\text{tossed})} = \frac{undefined}{undefined}.$$

Nor—perhaps, less importantly—would we be able to make judgments like the following, about a point (of area 0!) on the Earth's surface:

$$pr(\text{in western hemisphere} \mid \text{on equator}) = 1/2$$

even though

$$\frac{pr(\text{in western hemisphere} \wedge \text{on equator})}{pr(\text{on equator})} = \frac{0}{0}.$$

The quotient rule merely restates the product rule; and the product rule is no definition but an essential principle relating two distinct sorts of probability.

By applying the product rule to the terms on the right-hand sides of the analysis rules in sec. 1.3 we get the rule of[6]

Total Probability: If the D's partition $\top$
then $pr(H) = \sum_i pr(D_i)pr(H|D_i)$.[7]

Example. A ball will be drawn blindly from urn 1 or urn 2, with odds 2:1 of being drawn from urn 2. Is black or white the more probable outcome?



Solution. By the rule of total probability with H = black and D_i = drawn from urn i, we have $pr(H) = pr(H|D_1)P(D_1) + pr(H|D_2)P(D_2) = (\frac{3}{3} \cdot \frac{1}{3}) + (\frac{1}{2} \cdot \frac{2}{3}) = \frac{1}{4} \cdot \frac{1}{3} = \frac{7}{12} > \frac{1}{2}$: Black is the more probable outcome.

[6]Here the sequence of D's is finite or countably infinite.

[7]$= pr(D_1)pr(H|D_1) + pr(D_2)pr(H|D_2) + \ldots$.

1.5 Why '|' Cannot be a Connective

The bar in '$pr(H|D)$' is not a connective that turns pairs H, D of propositions into new, conditional propositions, *H if D*. Rather, it is as if we wrote the conditional probability of H given D as '$pr(H, D)$': the bar is a typographical variant of the comma. Thus we use 'pr' for a function of one variable as in '$pr(D)$' and '$pr(H \wedge D)$', and also for the corresponding function of two variables as in '$pr(H|D)$'. Of course the two are connected—by the product rule.

Then in fact we do not treat the bar as a statement–forming connective, 'if'; but why couldn't we? What would go wrong if we did? This question was answered by David Lewis in 1976, pretty much as follows.[8] Consider the simplest special case of the rule of total probability:

$$pr(H) = pr(H|D)pr(D) + pr(H|\neg D)pr(\neg D)$$

Now if '|' is a connective and D and C are propositions, then $D|C$ is a proposition too, and we are entitled to set $H = D|C$ in the rule of total probability. Result:

(1) $pr(D|C) = pr[(D|C)|D]pr(D) + pr[(D|C)|\neg D]pr(\neg D)$

So far, so good. But remember: '|' means *if*. Therefore, '$(D|C)|X$' means *If X, then if C then D*. And as we ordinarily use the word 'if', this comes to the same as *If X and C, then D*:

(2) $(D|C)|X = D|XC$

(Recall that the identity means the two sides represent the same region, i.e., the two sentences are logically equivalent.) Now by two applications of (2) to (1) we have

(3) $pr(D|C) = pr(D|D \wedge C)pr(D) + pr(D|\neg D \wedge C)pr(\neg D)$

But as $D \wedge C$ and $\neg(D \wedge C)$ respectively imply and contradict D, we have $pr(D|D \wedge C) = 1$ and $pr(D|\neg D \wedge C)) = 0$. Therefore, (3) reduces to

(4) $pr(D|C) = pr(D)$

[8] Lewis's "trivialization" result appeared in 'Probabilities of Conditionals and Conditional Probabilities' (1976), reprinted in his *Philosophical Papers, II*: Oxford University Press, 1986. For subsequent developments, see *Probabilities and Conditionals,* Ellery Eells and Brian Skyrms (eds.): Cambridge University press, 1994—especially, the papers by Alan Hàjek and Ned Hall.

Conclusion: If '|' were a connective ('if'), conditional probabilities would not depend on on their conditions at all. That means that '$pr(D|C)$' would be just a clumsy way of writing '$pr(D)$'. And it means that $pr(D|C)$ would come to the same thing as $pr(D|\neg C)$, and as $pr(D|X)$ for any other statement X.

That is David Lewis's "trivialization result". In proving it, the only assumption needed about 'if' was the eqivalence (2) of 'If X, then if C then D' with 'If X and C, then D'.[9]

1.6 Bayes's Theorem

A well-known corollary of the product rule allows us to reverse the arguments of the conditional probability function $pr(\ |\ )$ provided we multiply the result by the ratio of the probabilities of those arguments in the original order.

Bayes's Theorem (Probabilities). $pr(H|D) = pr(D|H) \times \dfrac{pr(H)}{pr(D)}$

Proof. By the product rule, the right-hand side equals $pr(D \wedge H)/pr(D)$; by the quotient rule, so does the left-hand side.

For many purposes Bayes's theorem is more usefully applied to odds than to probabilities. In particular, suppose there are two hypotheses, H and G, to which observational data D are relevant. If we apply Bayes theorem for probabilities to the conditional *odds* between H and G, $pr(H|D)/pr(G|D)$, the unconditional probability of D cancels out:

Bayes's Theorem (Odds). $\dfrac{pr(H|D)}{pr(G|D)} = \dfrac{pr(H)}{pr(G)} \times \dfrac{pr(D|H)}{pr(D|G)}$

Terminological note. The second factor on the right-hand side of the odds form of Bayes's theorem is the "LIKELIHOOD RATIO." In these terms the odds form says:

Conditional odds = prior odds × likelihood ratio

[9]Note that the result does not depend on assuming that 'if' means 'or not'; no such fancy argument is needed in order to show that $pr(\neg A \vee B) = pr(B|A)$ only under very special conditions. (Prove it!)

Bayes's theorem is often stated in a form attuned to cases in which you have clear probabilities $pr(H_1), pr(H_2), \ldots$ for mutually incompatible, collectively exhaustive hypotheses $H_1, H_2, \ldots$, and have clear conditional probabilities $pr(D|H_1)$, $pr(D|H_2)$, $\ldots$ for data D on each of them. For a countable collection of such hypotheses we have an expression for the probability, given D, of any one of them, say, H_i:

Bayes's Theorem (Total Probabilities):[10]

$$pr(H_i|D) = \frac{pr(H_i)pr(D|H_i)}{\sum_j pr(H_j)pr(D|H_j)}$$

EXAMPLE. In the urn example (**1.4**), suppose a black ball is drawn. Was it more probably drawn from urn 1 or urn 2? Let D = A black ball is drawn, H_i = It came from urn i, and $pr(H_i) = \frac{1}{2}$. Here are two ways to go:

(1) Bayes's theorem for total probabilities, $pr(H_1|D) = \frac{\frac{1}{2} \cdot \frac{3}{4}}{\frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{2}} = \frac{3}{5}$.

(2) Bayes's theorem for odds, $\frac{pr(H_1|D)}{pr(H_2|D)} = \frac{\frac{1}{2}}{\frac{1}{2}} \times \frac{\frac{3}{4}}{\frac{2}{4}} = \frac{3}{2}$.

These come to the same thing: Probability 3/5 = odds 3:2. Urn 1 is the more probable source.

1.7 Independence

Definitions

- $H_1 \wedge \ldots \wedge H_n$ is a **conjunction.** The H's are conjuncts.
- $H_1 \vee \ldots \vee H_n$ is a **disjunction.** The H's are disjuncts.
- For you, hypotheses are:

independent iff your probability for the conjunction of any two or more is the product of your probabilities for the conjuncts.

conditionally independent given G iff your conditional probability given G for the conjunction of any two or more is the product of your conditional probabilities given G for the conjuncts.

equiprobable (or conditionally equiprobable given G) iff they have the same probability for you (or the same conditional probability, given G).

[10] The denominator $= pr(H_1)pr(D|H_1) + pr(H_2)pr(D|H_2) + \cdots$.

EXAMPLE 1, **Rolling a fair die.** H_i means that the 1 ("ace") turns up on roll number i. These H_i are both independent and equiprobable for you: Your pr for the conjunction of any distinct n of them will be $1/6^n$.

EXAMPLE 2, **Coin and die**, $pr(\text{head}) = \frac{1}{2}$, $pr(1) = \frac{1}{6}$, $pr(\text{head} \wedge 1) = \frac{1}{12}$; outcomes of a toss and a roll are independent but not equiprobable.

EXAMPLE 3, **Urn of known composition.** You know it contains N balls, of which b are black, and that after a ball is drawn it is replaced and the contents of the urn mixed. $H_1, H_2, \ldots$ mean: the first, second, ... balls drawn will be black. Here, if you think nothing fishy is going on, you will regard the H's as equiprobable and independent: $pr(H_i) = b/N, pr(H_iH_j) = b^2/N^2$ if $i \neq j$, $pr(H_iH_jH_k) = b^3/N^3$ if $i \neq j \neq k \neq i$, and so on.

It turns out that three propositions H_1, H_2, H_3 can be independent in pairs but fail to be independent because $pr(H_1 \wedge H_2 \wedge H_3) \neq \text{pr}(\text{H}_1)pr(H_2)pr(H_3)$.

EXAMPLE 4. **Two normal tosses of a normal coin.** If we define H_i as 'head on toss $\#i$' and D as 'different results on the two tosses' we find that $pr(H_1 \wedge H_2) = pr(H_1)pr(D) = pr(H_2)pr(D) = \frac{1}{4}$ but $pr(H_1 \wedge H_2 \wedge D) = 0$.

Here is a useful fact about independence:

> (1) If n propositions are independent, so are those obtained by denying some or all of them.

To illustrate (1), think about case $n = 3$. Suppose H_1, H_2, H_3 are independent. Writing $h_i = pr(H_i)$, this means that the following four equations hold:

$$\text{(a) } pr(H_1 \wedge H_2 \wedge H_3) = h_1h_2h_3,$$

$$\text{(b) } pr(H_1 \wedge H_2) = h_1h_2, \text{ (c) } pr(H_1 \wedge H_3) = h_1h_3, \text{ (d) } pr(H_2 \wedge H_3) = h_2h_3$$

Here, $H_1, H_2, \neg H_3$ are also independent. Writing $\bar{h}_i = pr(\neg H_i)$, this means:

$$\text{(e) } pr(H_1 \wedge H_2 \wedge \neg H_3) = h_1h_2\bar{h}_3,$$

$$\text{(f) } pr(H_1 \wedge H_2) = h_1h_2, \text{ (g) } pr(H_1 \wedge \neg H_3) = h_1\bar{h}_3, \text{ (h) } pr(H_2 \wedge \neg H_3) = h_2\bar{h}_3$$

Equations $(e) - (h)$ follow from $(a) - (d)$ by the rule for 'but not.'

EXAMPLE 5, **(h) follows from (b).** $pr(H_2 \wedge \neg H_3) = h_2\bar{h}_3$ since, by 'but not', the left-hand side $= pr(H_1) - pr(H_1 \wedge H_2)$, which, by (b), $= h_1 - h_1h_2$, and this $= h_1(1 - h_2) = h_1\bar{h}_2$.

A second useful fact follows immediately from the quotient rule:

> (2) If $pr(H_1) > 0$, then H_1 and H_2 are independent iff $pr(H_2|H_1) = pr(H_2)$.

1.8 Exchangeability[11]

In this section we consider propositions $H_1, \ldots, H_n$ which are **exchangeable** for you in the sense that your probability for some being true and the others false depends only on the *numbers* supposed to be true and false. Formally:

> For each $t = 0, \ldots, n$, you assign the same probability to all conjunctions of t of $H_1, \ldots, H_n$ with the denials of the others.

EXAMPLE 1, $\boldsymbol{n = 3}$. If we write 'H_i' for 'head on the i'th toss' then for D = 'same outcome on the first two tosses' we write $(H_1 \wedge H_2) \vee (\neg H_1 \neg H_2)$, abbreviated $H_1H_2 \vee \bar{H}_1\bar{H}_2$. Here, between the double lines, is the table for D as a truth-function of H_1, H_2, H_3, with '1' for $true$ and '0' for $false$:

		H_1	H_2	H_3	D	Your $pr(H_t^3)$'s	Your $pr(C_k)$'s
H_3^3	C_1	1	1	1	1	.2	.2
	C_2	1	1	0	1		.1
H_2^3	C_3	1	0	1	0	.3	.1
	C_4	0	1	1	0		.1
	C_5	1	0	0	0		.1
H_1^3	C_6	0	1	0	0	.3	.1
	C_7	0	0	1	1		.1
H_0^3	C_8	0	0	0	1	.2	.2

Notation Truth table for D One probability assignment

To the left of the double lines, each C_k is a conjunction describing the assignment to its right of truth values to the H_i. Thus $C_1 = H_1H_2H_3$ and $C_2 = H_1H_2\bar{H}_3$. The **full normal form** of D is the disjunction $C_1 \vee C_2 \vee C_7 \vee C_8$ of the 3-termed conjunctions describing D's four '1' cases. To the right of the truth table is one probability assignment that makes the H_i exchangeable.

In general, we write 'H_t^n' for the statement that the total number of truths among the n H_i is exactly t, as in $H_3^3 = C_1$ and $H_2^3 = C_2 \vee C_3 \vee C_4$:

> H_t^n = the disjunction of all $\binom{n}{t} = \frac{n!}{t!f!}$ conjunctions of t of $H_1, \ldots, H_n$ with the denials of the other f $(= n - t)$.

[11] Bruno de Finetti, "La prévision, ses lois logiques, ses sources subjectives," *Annales de l'institut Henri Poincaré,* **7**(1937)1-68; chapter III. Translation ("Foresight . . . ") in *Studies in Subjective Probability,"* Henry Kyburg, Jr. and Howard Smokler (eds.), Krieger, New York (1980) 53-118.

Now the definition of exchangeability implies that

(1) $pr(K) = \frac{pr(H_t^n)}{\binom{n}{t}}$, where K is the conjunction of t of $H_1, \ldots, H_n$ with the denials of the other $n - t$.

It follows that exchangeability undoes the dreaded "combinatorial explosion" by allowing your probabilities for all truthfunctional compounds of n exchangeable H_i to be determined by n of those compounds, when in general it takes $2^n - 1$ to do the job (for example, if $n = 3$, all but one of the C_k):

Implosion: If you see $H_1, \ldots, H_n$ as exchangeable, your probabilities for all $2^{(2^n)}$ of their distinct truthfunctional compounds are determined by your probabilities for any n of $H_0^n, \ldots, H_n^n$.

Proof. By additivity, $pr(D) = \sum_K pr(K)$ for the K's that appear as conjuncts in D's full normal form. By (1) these $pr(K)$'s are determined by your $pr(H_t^n)$'s. Since $pr(H_0^n) + \ldots + pr(H_n^n) = 1$, any n of these terms determine all $n + 1$.

Problem 20 in sec. **1.9** is a walk through a proof of the following important corollary of the combinatorial implosion.[12]

The Rule of Succession: If $H_1, \ldots, H_{n+1}$ are exchangeable for you, your probability for H_{n+1} conditionally on a conjunction K of t of $H_1, \ldots, H_n$ with the denials of the $f\ (= n - t)$ others is

$$pr(H_{n+1}|K) = pr(H_{n+1}|H_t^n) = \frac{t+1}{n+2+*},$$

where the correction term, $*$, is $(f+1)(\frac{pr(H_t^{n+1})}{pr(H_{t+1}^{n+1})} - 1)$.

EXAMPLE 2, **Uniform distribution:** $pr(H_t^{n+1}) = \frac{1}{n+2}$ for $t = 0, \ldots, n+1$. In this ("Bayes-Laplace-Johnson-Carnap") case the correction term vanishes, and the rule of succession reduces to

$$pr(H_{n+1}|K) = pr(H_{n+1}|H_t^n) = \frac{t+1}{n+2},$$

where t is the number of undenied H_i in K.

EXAMPLE 3, **An urn of unknown composition.** Normally, if you don't know the proportion p of black balls in an urn you will regard the propositions

[12] The best overall account is by Sandy Zabell, "The Rule of Succession" *Erkenntnis* **31** (1989) 283-321.

H_i that black is drawn on trial $i = 1, \ldots, n$ as exchangeable. If, in particular, you adopt the uniform distribution, your probability for black next, given 6 black on the first 8 trials, will be $pr(H_9|H_6^8) = pr(H_9|K)$, where K is any disjunct of H_6^8. Then your $pr(H_9|H_6^8) = \dfrac{t+1}{n+2} = \dfrac{7}{10}$.

1.9 Problems

1 Find formulas for
(1) $pr(F \vee G \vee H)$ and
(2) $pr(H_1 \vee \ldots \vee H_n)$
in terms of probabilities of some of $H_1, \ldots, H_n$ and their conjunctions.

2 Exclusive 'or'. The symbol "$\vee$" stands for "or" in a sense that is not meant to rule out the possibility that the hypotheses flanking it are both true: $G \vee H = G \vee H \vee (G \wedge H)$. Let us use symbol $\underline{\vee}$ for "or" in an exclusive sense: $G \underline{\vee} H = (G \vee H) \wedge \neg(G \wedge H)$. Find formulas for
(1) $pr(G \underline{\vee} H)$ and
(2) $pr(F \underline{\vee} G \underline{\vee} H)$
in terms of probabilities of some of $H_1, \ldots, H_n$ and their conjunctions.
(3) Does $F \underline{\vee} G \underline{\vee} H$ mean that exactly one of F, G, H is true? (No.) What *does* it mean?
(4) What does $H_1 \underline{\vee} \ldots \underline{\vee} H_n$ mean?

3 *Diagnosis.*[13] The patient has a breast mass that her physician thinks is probably benign: frequency of malignancy among women of that age, with the same symptoms, family history, and physical findings, is about 1 in 100. The physician orders a mammogram and receives the report that in the radiologist's opinion the lesion is malignant, i.e., the mammogram is positive. Based on the available statistics, the physician's probabilities for true and false positive mammogram results were as follows, and her prior probability for the patient's having cancer was 1%. What will her conditional odds on malignancy be, given the positive mammogram?

$pr(row\|column)$	Malignant	Benign
+ mammogram	80%	10%
− mammogram	20%	90%

[13] Adapted from David M. Eddy's 'Probabilistic reasoning in clinical medicine', which appears in a useful collection of articles edited by Daniel Kahneman, Paul Slovic, and Amos Tversky: *Judgment Under Uncertainty* (Cambridge University Press, 1982).

4 *The Taxicab Problem.*[14] "A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:
"(a) 85% of the cabs in the city are Green, 15% are Blue.
"(b) A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. What is the probability that the cab involved in the accident was Blue rather than Green [i.e., conditionally on the witness's identification]?"

5 *The Device of Imaginary Results.*[15] This is meant to help you identify your prior odds—e.g., on the hypothesis "that a man is capable of extrasensory perception, in the form of telepathy. You may imagine an experiment performed in which the man guesses 20 digits (between 0 and 9) correctly. If you feel that this would cause the probability that the man has telepathic powers to become greater than 1/2, then the [prior odds] must be assumed to be greater than 10^{-20}. ... Similarly, if three consecutive correct guesses would leave the probability below 1/2, then the [prior odds] must be less than 10^{-3}."
Verify these claims about the prior odds.

6 *The Rare Disease.*[16] "You are suffering from a disease that, according to your manifest symptoms, is either A or B. For a variety of demographic reasons disease A happens to be 19 times as common as B. The two diseases are equally fatal if untreated, but it is dangerous to combine the respective appropriate treatments. Your physician orders a certain test which, through the operation of a fairly well understood causal process, always gives a unique diagnosis in such cases, and this diagnosis has been tried out on equal numbers of A and B patients and is known to be correct on 80% of those occasions. The tests report that you are suffering from disease B. Should you nevertheless opt for the treatment appropriate to A, on the supposition that the probability of your suffering from A is 19/23? Or should you opt for the treatment appropriate to B, on the supposition [...] that the probability of your suffering from B is 4/5? It is the former opinion that would be irrational for you. Indeed, on the other view, which is the one espoused in the literature, it would be a waste of time and money even to carry out the tests, since whatever their results, the base rates would still compel a more than 4/5 probability in favor of disease A. So the literature is propagating an analysis

[14]From *Judgment Under Uncertainty,* pp. 156-158.
[15]From I. J. Good, *Probability and the Weighing of Evidence*: London, 1950, p. 35.
[16]L. J. Cohen, in *The Behavioral and Brain Sciences* **4**(1981)317-331; see p. 329.

that could increase the number of deaths from a rare disease of this kind." Diaconis and Freedman (1981, pp. 333-4) suggest that "the fallacy of the transposed conditional" is being committed here, i.e., confusion of the following quantities—the second of which is the true positive rate of the test for B: $pr(It\ is\ B|It\ is\ diagnosed\ as\ B)$, $pr(It\ is\ diagnosed\ as\ B|It\ is\ B)$.
Use the odds form of Bayes's theorem to verify that if your prior odds on A are 19:1 and you take the true positive rate (for A, and for B) to be 80%, your posterior probability for A should be 19/23.

7 *On the Credibility of Extraordinary Stories.*[17]
"There are, broadly speaking, two different ways in which we may suppose testimony to be given. It may, in the first place, take the form of a reply to an alternative question, a question, that is, framed to be answered by yes or no. Here, of course, the possible answers are mutually contradictory, so that if one of them is not correct the other must be so: —Has A happened, yes or no?" ...
"On the other hand, the testimony may take the form of a more original statement or piece of information. Instead of saying, Did A happen? we may ask, What happened? Here if the witness speaks the truth he must be supposed, as before, to have but one way of doing so; for the occurrence of some specific event was of course contemplated. But if he errs he has many ways of going wrong" ...
(a) In an urn with 1000 balls, one is green and the rest are red. A ball is drawn at random and seen by no one but a slightly colorblind witness, who reports that the ball was green. What is your probability that the witness was right on this occasion, if his reliability in distinguishing red from green is .9, i.e., if $pr(He\ says\ it\ is\ X|\text{It is } X) = .9$ when $X =$ Red and when $X =$ Green?
(b) "We will now take the case in which the witness has many ways of going wrong, instead of merely one. Suppose that the balls were all numbered, from 1 to 1000, and the witness knows this fact. A ball is drawn, and he tells me that it was numbered 25, what is the probability that he is right?" In answering you are to "assume that, there being no apparent reason why he should choose one number rather than another, he will be likely to announce all the wrong ones equally often."
What is now your probability that the 90% reliable witness was right?

8.1 *The Three Cards.* One is red on both sides, one is black on both sides, and the other is red on one side and black on the other. One card is drawn blindly and placed on a table. If a red side is up, what is the probability that

[17] Adapted from pp. 409 ff. of John Venn, *The Logic of Chance,* 3rd. ed, 1888 (Chelsea Pub. Co. reprint, 1962).

the other side is red too?

8.2 *The Three Prisoners.* An unknown two will be shot, the other freed. Alice asks the warder for the name of one other than herself who will be shot, explaining that as there must be at least one, the warder won't really be giving anything away. The warder agrees, and says that Bill will be shot. This cheers Alice up a little: Her judgmental probability for being shot is now 1/2 instead of 2/3. Show (via Bayes's theorem) that
(a) Alice is mistaken if she thinks the warder is as likely to say 'Clara' as 'Bill' when he can honestly say either; but that
(b) She is right if she thinks the warder will say 'Bill' when he honestly can.

8.3 *Monty Hall.* As a contestant on a TV game show, you are invited to choose any one of three doors and receive as a prize whatever lies behind it—i.e., in one case, a car, or, in the other two, a goat. When you have chosen, the host opens one of the other two doors, behind which he knows there is a goat, and offers to let you switch your choice to the third door. Would that be wise?

9 *Causation vs. Diagnosis.*[18] "Let A be the event that before the end of next year, Peter will have installed a burglar alarm in his home. Let B denote the event that Peter's home will have been burgled before the end of next year.

"Question: Which of the two conditional probabilities, $pr(A|B)$ or $pr(A|\neg B)$, is higher?
"Question: Which of the two conditional probabilities, $pr(BA)$ or $pr(B|\neg A)$, is higher?
"A large majority of subjects (132 of 161) stated that $pr(A|B) > pr(A|\neg B)$ and that $pr(B|A) < pr(B|\neg A)$, contrary to the laws of probability."
Substantiate this critical remark by showing that the following is a law of probability.
$pr(A|B) > pr(A|\neg B)$ iff $pr(B|A) > pr(B|\neg A)$

10 Prove the following, assuming that conditions all have probability > 0.

a If A implies D then $pr(A|D) = pr(A)/pr(D)$.

b If D implies A then $pr(\neg A|\neg D) = pr(\neg A)/pr(\neg D)$. (*"TJ's Lemma"*)

c $pr(C|A\vee B)$ is in the interval from $pr(C|A)$ to $pr(C|B)$ if $pr(A\wedge B) = 0$.

11 *Sex Bias at Berkeley?*[19] In the fall of 1973, when 8442 men and 4321 women applied to graduate departments at U. C. Berkeley, about 44% of

[18] From p. 123 of *Judgment Under Uncertainty.*

[19] David Freedman, Robert Pisani, and Roger Purves, *Statistics*, W. W. Norton, New York, 1978, pp. 12-15.

the men were admitted, but only about 35% of the women. It looked like sex bias against women. But when admissions were tabulated for the separate departments—as below, for the six most popular departments, which together accounted for over a third of all the applicants—there seemed to be no such bias on a department-by-department basis. And the tabulation suggested an innocent one-sentence explanation of the overall statistics. What was it?
Hint: What do the statistics indicate about how hard the different departments were to get into?

Department A admitted 62% of 825 male applicants, 82% of 108 females.
Department B admitted 63% of 560 male applicants, 68% of 25 females.
Department C admitted 37% of 325 male applicants, 34% of 593 females.
Department D admitted 33% of 417 male applicants, 35% of 375 females.
Department E admitted 28% of 191 male applicants, 24% of 393 females.
Department F admitted 6% of 373 male applicants, 7% of 341 females.

12 *The Birthday Problem.* Of twenty-three people selected at random, what is the probability that at least two have the same birthday?

$$Hint: \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \cdots (23 factors) \approx .49$$

13 *How would you explain the situation to Méré?*
"M. de Méré told me he had found a fallacy in the numbers for the following reason: With one die, the odds on throwing a six in four tries are 671:625. With two dice, the odds are against throwing a double six in four tries. Yet 24 is to 36 (which is the number of pairings of the faces of two dice) as 4 is to 6 (which is the number of faces of one die). This is what made him so indignant and made him tell everybody that the propositions were inconsistent and that arithmetic was self-contradictory; but with your understanding of the principles, you will easily see that what I say is right." (Pascal to Fermat, 29 July 1654)

14 *Independence, sec.* **1.7**.
(a) Complete the proof, begun in example 5, of the case $n = 3$ of fact (1).
(b) Prove fact (1) in the general case. Suggestion: Use mathematical induction on the number $f = 0, 1, \ldots$ of denied H's, where $2 \leq n = t + f$.

15 *Exchangeability.* Show that
(a) Independence implies exchangeability.
(b) If $H_1, \ldots, H_n, H_{n+1}$ are exchangeable, so are $H_1, \ldots, H_n$.

16 *A Pólya Urn Model.* At stage 0, an urn contains two balls, one black and one white. At each later stage a ball is drawn with *double replacement:* it is replaced along with another ball of its color. Prove that the conditional

odds on black's being drawn on the $n+1$'st trial, given that black was drawn on the first t trials and white on the next $f = n - t$, are $\frac{t+1}{f+1}$, exactly as in EXAMPLE 2 of sec. **1.8**.

17 *Exchangeability is preserved under mixing.* Prove that if $H_1, \ldots, H_n$ are exchangeable relative to each of $pr_1, \ldots, pr_m$ then they are exchangeable relative to $w_1 pr_1 + \ldots + w_m pr_m$ if the w's are non-negative and sum to 1.

18 *Independence is not preserved under mixing.* Prove this, using a simple counterexample where $m = 2$ and $w_1 = w_2$.

19 *Independence, Conditional Independence, Exchangeability.* Suppose you are sure that the balls in an urn are either 90% red or 10% red, and you regard those two hypotheses as equiprobable. If balls are drawn with replacement, then conditionally on each hypothesis the propositions $H_1, H_2, \ldots$ that the first, second, etc., balls drawn will be red are independent and equiprobable. Is that also so unconditionally?
(a) Are the R's unconditionally equiprobable? If so, what is their probability? If not, why not?
(b) Are they unconditionally independent? Would your probability for red next be the same as it was initially after drawing one ball, which is is red?
(c) Are they are exchangeable?

20 *Proof of the rule of succession in sec.* **1.8.** Explain steps (b)-(g):

(a) $pr(H_{n+1}|K) = \dfrac{pr(H_{n+1} \wedge K)}{pr(K)} = \dfrac{pr(H_{t+1}^{n+1})/\binom{n+1}{t+1}}{pr(H_t^n)/\binom{n}{t}} = \dfrac{(t+1)\,pr(H_{t+1}^{n+1})}{(n+1)\,pr(H_t^n)}$ by the quotient rule, implosion, and the fact that $\binom{n}{t}/\binom{n+1}{t+1} = (t+1)/(n+1)$.

(b) $pr(K) = pr(\neg H_{n+1} \wedge K) + pr(H_{n+1} \wedge K)$

(c) $\dfrac{pr(H_t^n)}{\binom{n}{t}} = \dfrac{pr(H_t^{n+1})}{\binom{n+1}{t}} + \dfrac{pr(H_{t+1}^{n+1})}{\binom{n+1}{t+1}}$

(d) $\dfrac{\binom{n}{t}}{\binom{n+1}{t}} = \dfrac{f+1}{n+1}$ and $\dfrac{\binom{n}{t}}{\binom{n+1}{t+1}} = \dfrac{t+1}{n+1}$

(e) $pr(H_t^n) = \dfrac{f+1}{n+1} pr(H_t^{n+1}) + \dfrac{t+1}{n+1} pr(H_{t+1}^{n+1})$

(f) $\dfrac{(t+1)\,pr(H_{t+1}^{n+1})}{(n+1)\,pr(H_t^n)} = \dfrac{(t+1)pr(H_{t+1}^{n+1})}{(f+1)pr(H_t^{n+1}) + (t+1)(\frac{pr(H_t^{n+1})}{pr(H_{t+1}^{n+1})} - 1)}$

(g) The right-hand side of (f) $= \dfrac{t+1}{n+2+*}$.

21 *Sampling with double replacement:* Pólya's urn model. Initially an urn contains 2 balls, one black and one red. On each trial a ball is drawn at

random, and replaced along with another ball of its color.

(a) Show that the outcomes are exchangeable.

(b) What is the rule of succession in this case?

(c) What is the rule of succession when the urn contains b black and r red balls initially, and sampling is with m-tuple replacement?

Chapter 2

Dynamics; Learning

So far, we have been concerned with your probabilities at a particular time, as represented by single probability assignment, *pr*. But now we turn to the matter of learning from experience, where a pair of probability assignments represent your judgments before and after you change your mind in response to the result of some experiment or observation. We start with the simplest, most familiar mode of updating—which will then be generalized, in sec. **2.3**.

2.1 Conditioning

Suppose your judgments were once characterized by an assignment ("*old*") which describes all of your conditional and unconditional probabilities as they were then. And suppose you became newly convinced of the truth of some data statement D. That would change your probabilistic judgments; they would no longer corrrespond to your prior assignment *old*, but to some posterior assignment *new*, where

CERTAINTY:
$$new(D) = 1$$

How should *new* be related to *old*? The simplest answer goes by the name "conditioning" (or "conditionalization") on D:

CONDITIONING:
$$new(-) = old(-|D)$$

This means that your new unconditional probability for any hypothesis H will simply be your old conditional probability for H given D.

When is it appropriate to update by conditioning on D? It is easy to see—once you think of it—that certainty about D is not enough.

EXAMPLE 1, **The Red Queen**. You learn that a playing card, drawn from a well-shuffled, standard 52-card deck, is a heart: your $new(\heartsuit) = 1$. Since you know that all hearts are red, this means that your $new(\text{red}) = 1$ as well. But since $old(\text{Queen} \wedge \heartsuit|\heartsuit) = \frac{1}{13}$ whereas $old(\text{Queen} \wedge \heartsuit|\text{red}) = \frac{1}{26}$, you will have different new probabilities for the card's being the Queen of hearts, depending on which of those certainties you update on.

Of course, in example 1 we are in no doubt about which certainty you should condition on: it is the more informative one, $\heartsuit$. And if what you actually learned was that the card was a court card as well as a heart, then that is what you ought to update on, so that your $new(Q \wedge \heartsuit) = old(Q \wedge \heartsuit|(\text{Ace} \vee \text{King} \vee \text{Queen}) \wedge \heartsuit)) = \frac{1}{3}$.

But what if—although the thing you learned with certainty was indeed that the card was a heart—you also saw something you can't put into words, that gave you reason to suspect, uncertainly, that it was a court card.

EXAMPLE 2, **A brief glimpse** determines your new probabilities as 1/20 for each of the ten numbered hearts, and as 1/6 for each of the three others. Here, where some of what you have learned is just probabilistic, conditioning on your most informative new certainty does not take all of your information into account. In sec. **2.3** we shall see how you might take such uncertain information into account; but meanwhile, it is clear that ordinary conditioning is not always the way to go.

When *does* it make sense for you to plan to update by conditioning on D if you should become sure that D is true? One obvious answer is "rigidity" of conditional probabilities: confidence that your conditional probabilities given D will not be changed by an experience that makes you certain of D's truth.

RIGIDITY:
$$new(-|D) = old(-|D)$$

It is easy to see that certainty and rigidity together imply conditioning:[1]

[1]Proof: $old(H|D) = new(H|D)$ by (b), which $= new(H)$ by (a).

Premise (a): $new(D) = 1$

Premise (b): $new(H|D) = old(H|D)$

Conclusion: $new(H) = old(H|D)$

And conversely, since both of the following inferences are valid, conditioning implies both certainty and rigidity.[2]

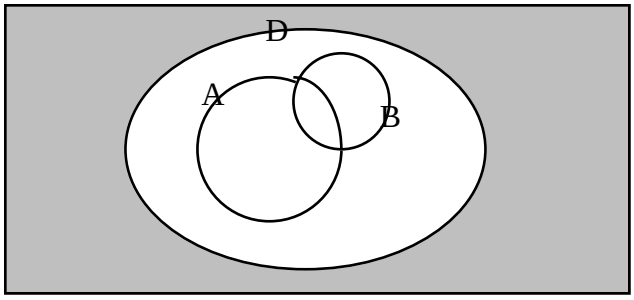
Premise: $new(H) = old(H|D)$

Conclusion (a): $new(D) = 1$

Conclusion (b): $new(H|D) = old(H|D)$

Rigidity is invariance of conditional probabilities given D. Another intersting invariance is that of ratios of unconditional probabilities of different ways in which D might be true,

PROPORTIONALITY: $\dfrac{new(A)}{new(B)} = \dfrac{old(A)}{old(B)}$ if A and B each imply D.



It is easy to see that proportionality implies, and is implied by, yet another invariance,[3]

BASIC PROPORTIONALITY: $\dfrac{new(A)}{new(D)} = \dfrac{old(A)}{old(D)}$ if A implies D.

[2]Proof of (a): Setting $H = D$ in the premise, we have $new(D) = old(D|D)$, which =1 by the quotient rule. Proof of (b): By (a), $new(H) = new(H \wedge D) = \dfrac{new(H \wedge D)}{new(D)}$, which $= new(H|D)$ by the quotient rule.

[3]Proof. Implies: set $B = D$ in proportionality. Is implied by: replace A by B in basic proportionality and compare the two versions of $new(D)/old(D)$.

Now in fact, all three invariances are equivalent. (Easy proof: Show that rigidity implies and is implied by basic proportionality.)

There are special circumstances in which the rigidity condition dependably holds:

EXAMPLE 3, **The Statistician's Stooge.**[4] Rigidity will surely hold in the variant of example 1 in which you do not see the card yourself, but have arranged for a trusted assistant to look at it and tell you simply whether or not is is a heart, with no further information. Under such conditions your unconditional probability that the card is a heart changes to 1 in a way that cannot change any of your conditional probabilities, given ♡.

2.2 Probabilities from Statistics

We can think of observations and experiments as ways of putting multiple-choice questions to Nature. In the statistician's stooge example, if you trust your assistant to tell you the suit of the card, and only that, there are four possible answers, and the invariance condition will be satisfied relative to each: $new(H|\heartsuit) = old(H|\heartsuit)$, $new(H|\diamondsuit) = old(H|\diamondsuit)$, $new(H|\spadesuit) = old(H|\spadesuit)$, $new(H|\clubsuit) = old(H|\clubsuit)$. If your assistant says '♡', the certainty condition holds in the form $new(\heartsuit) = 1$, $new(\diamondsuit) = new(\spadesuit) = new(\clubsuit) = 0$.

In the examples to which we now turn, the familiar "frequentist" account of how probabilities ought to be formed is shown to be correct under a special assumption that is often found compelling.

EXAMPLE 1, **Tossing a bent coin.** Here the question put to Nature is: 'What is the fraction s/n of successes (say, heads) among the n trials of an experiment (say, tossing a bent coin) that have been made so far?' There are $n + 1$ possible answers; in the notation of sec. **1.8** these are H_t^n with $t = 0, 1, \ldots, n$.

[4] I. J. Good's term.

2.3 Generalized Conditioning[5]

Certainty is quite demanding. It rules out not only the far-fetched uncertainties associated with philosophical skepticism, but also the familiar uncertainies that affect real empirical inquiry in science and everyday life. But it is a condition that turns out to be dispensible: As long as rigidity holds, updating is valid by a generalization of conditioning to which we now turn. We begin with the simplest case, of a yes/no question.

If rigidity holds relative to each answer to a yes/no question, and something makes you (say) 85% sure that the answer is 'yes', a way of updating your probabilities is still available for you. Rigidity means that $new(H|\text{yes}) = old(H|yes)$ and $new(H|\text{no}) = old(H|\text{no})$ for each hypothesis H. The required updating rule is easily obtained from the law of total probability in the form

$$new(H) = new(H|\text{yes})new(\text{yes}) + new(H|\text{no})new(\text{no}).$$

Rewriting the two conditional probabilities in this equation via the rigidity conditions, we have the following updating scheme:

$$new(H) = old(H|\text{yes})new(\text{yes}) + old(H|\text{no})new(\text{no})$$

More generally, with a countable partition of answers, the applicable rule of total probability has one term on the right for each member of the partition. If the rigidity condition holds for each answer, D_i, then we have the updating scheme $new(H) = old(H|D_1)new(D_1) + old(H|D_2)new(D_2) + \ldots$, called

PROBABILITY KINEMATICS:

$$new(H) = \sum_i old(H|D_i)new(D_i).$$

EXAMPLE, **An Expert's Observations.** In hopes of settling on one of the following diagnoses, Dr. Jane Doe, a histopathologist, will conduct a microscopic examination of a section of tissue surgically removed from a tumor. She is sure that exactly one of the three is correct.

D_1 = Islet cell carcinoma, D_2 = Ductal cell carcinoma, D_3 = Benign tumor.

Here $n = 3$, so $new(H) =$

[5]Also known as *probability kinematics* and *Jeffrey conditioning*. For a little more about this, see Persi Diaconis and Sandy Zabell, 'Some alternatives to Bayes's rule' in *Information Pooling and Group Decision Making,* Bernard Grofman and Guillermo Owen (eds.), JAI Press, Greenwich, Conn. and London, England, pp. 25-38. For much more, see 'Updating subjective probability' by the same authors, *Journal of the American Statistical Association* **77**(1982)822-830.

$$old(H|D_1)new(D_1) + old(H|D_2)new(D_2) + old(H|D_3)new(D_3).$$

In the event, the examination does not drive her probability for any of the diagnoses to 1; rather, it leads her to assign them these probabilities:

$$new(D_1) = \tfrac{1}{3}, \;\; new(D_2) = \tfrac{1}{6}, \;\; new(D_3) = \tfrac{1}{2}.$$

But suppose her conditional probabilities for the hypothesis H of 5 year survival given the diagnoses are unaffected by her examination:

$$\begin{aligned} old(H|D_1) &= new(H|D_1) = .4, \\ old(H|D_2) &= new(H|D_2) = .6, \\ old(H|D_1) &= new(H|D_3) = .9. \end{aligned}$$

Then by generalized conditioning, her new probability for 5 year survival will be a weighted average

$$new(H) = (.4)(\tfrac{1}{3}) + (.6)(\tfrac{1}{6}) + (.9)(\tfrac{1}{2}) \approx .7$$

of the values that $old(H)$ would have had if she had been sure of the three diagnoses, where the weights are her new probabilities for those diagnoses.

NOTE: The calculation of $new(H)$ makes no use of Dr Doe's old probabilities for the diagnoses. As far as that goes, there might have been no $old(D_i)$'s at all that characterized her prior mental state. Nevertheless, her $new(D_i)$'s may well have arisen as a joint effect of her $old(D_i)$'s and her new experiences at the microscope. The kinematical formula uses the results, $new(D_i)$, of any such interaction as data; the formula itself does not compute those results.

2.4 "Real" Probability

It is natural to think there is such a thing as real or objective probability—"chance", for short—in contrast to merely judgmental probability.

EXAMPLE 1, **The Urn.** An urn contains 100 balls, of which an unknown number n are of the winning colorgreen, say. You regard the drawing as honest in the usual way. Then you think the chance of winning is n/100. If all 101 possible values of n are equiprobable in your judgment, then your prior probability for winning on any trial is 50your posterior odds between

any two possible compositions of the urn will change from 1:1 to various other values, even though you are sure that the composition itself doesnt change.

"The chance of winning is 30%." What does that mean? In the urn example, it means that $n = 30$. How can we find out whether it is true or false? In the urn example we just count the green balls. But in other casesdie rolling, horse racing, etc.there may be no process that does the Count the green ones job. In general, there are puzzling questions about the hypothesis that the chance of H is p that dont arise regarding the hypothesis H itself. David Humes skeptical answer to those questions says that chances are simply projections of robust features of judgmental probabilities from our minds out into the world, whence we hear them clamoring to be let back in. Thats how our knowledge that the chance of H is p guarantees that our judgmental probability for H is p: the guarantee is really a presupposition. As Hume sees it, the argument

(1) pr(the chance of H is p) $= 1$, so $pr(H) = p$.

is valid because our conviction that the chance of H is p is a just a firmly felt commitment to p as our judgmental probability for H. What if we are not sure what the chance of H is, but think it may be p? Here, the relevant principle ("Homecoming") specifies the probability of H given that its chance is pexcept in cases where we are antecedently sure that the chance is not p because, for some chunk $(\cdots\cdots)$ of the interval from 0 to 1, pr(the chance of H is inside the chunk) $= 0$.

> HOMECOMING. $pr(H|$ chance of H is $p) = p$ unless p is excluded as a possibile value, being in the interior of an interval we are sure does not contain the chance of H :
>
> 0——-$\cdots p \cdots$—1

Note that when pr(the chance of H is p) $= 1$, homecoming validates argument (1). The name 'Homecoming' is loaded with philosophical baggage. 'Decisiveness' would be a less tendentious name, acceptable to those who see chances as objective features of the world, independent of what we may think. The condition 'the chance of H is p' is decisive in that it overrides any other evidence represented in the probability function pr. But a decisive condition needn not override other conditions conjoined with it to the right of the bar. In particular, it will not override the hypothesis that H is true, or that H is false. Thus, since $pr(H|H \wedge C) = 1$ when C is any condition consistent with C, we have

(2) $pr(H|H \wedge$ the chance of H is .3) $= 1$, not .3,

and that is no violation of decisiveness. On the Humean view it is ordinary conditions, making no use of the word 'chance,' that appear to the right of the bar in applications of the homecoming principle, e.g., conditions specifying the composition of an urn. Here, you are sure that if you knew n, your judgmental probability for the next balls being green would be $n\%$:

(3) pr(Green next $|n$ of the hundred are green) $= n\%$

Then you take the chance of green next to be a physical magnitude, $n/100$, which you can determine empirically by counting. It is the fact that for you n satisfies the decisiveness condition that identifies $n\%$ as the chance of green next, in your judgment. Its easy in the urn example: an observable magnitude n/100 turns out to satisfy your decisiveness condition, and so was identifiable as your idea of the chance of Hs being true. Loaded dice are a harder problem.

EXAMPLE 2, **The Loaded Die.** Suppose H predicts ace on the next toss. Perhaps you are sure that if you understood the physics better, knowledge of the mass distribution in the die would determine for you a definite judgmental probability of ace next: if you knew the physics, then for some f, principle (4), below, would apply. But you don't know the physics; in this case there is no physical magnitude, f, that now satisfies the decisiveness condition for your judgmental probabilitiy function, pr.

(4) pr(Head next | The mass distribution is M) $= f(M)$

When we are unlucky in this way—when there is no decisive physical parameter for us—there may still be some point in speaking of the chance of H, i.e., of a yet-to-be-identified physical parameter that will be decisive for people in the future. In the decisiveness condition we might read 'chance of H is p' as a place-holder for some presently unavailable description of a presently unidentified physical parameter that we will recognize as decisive in the future. There is no harm in that–as long as we don't think we are already there.

In the urn example it is clear to us what physical parameter, X, plays the role of chance, and we can specify the function, f, that maps X into the chance: X is $n/100$ and f is the identity function, $f(X) = X$. In the die example we are clear about the parameter X, but not about the function f. And in other cases we are also unclear about X, identifying it in terms of its salient effect ("blood poinsoning" in the following example) while seeking clues about the cause.

EXAMPLE 3, **Blood Poisoning.**

> 'At last, early in 1847, an accident gave Semmelweis the decisive clue for his solution of the problem. A colleague of his,

> Kolletschka, received a puncture wound in the finger, from the scalpel of a student with whom he was performing an autopsy, and died after an agonizing illness during which he displayed the same symptoms that Semmelweis had observed in the victims of childbed fever. Although the role of microorganisms in such infections had not yet been recognized at that time, Semmelweis realized that "cadaveric matter" which the student's scalpel had introduced into Kolletschchka's blood stream had caused his colleague's fatal illness. And the similarities between the course of Kolletschchka's disease and that of the women in his clinic led Semmelweis to the conclusion that his patients had died of the same kind of blood poisoning"[6]

But what is the common character of the X's that we had in the urn and die examples, and that Semmelweis lacked, and of the f's that we had in the first, but lacked in the other two? These strike me as a questions in pragmatics, not semantics; they concern the place of these X's and f's in our processes of judgment. The answers have to do with invariance of conditional judgmental probabilities as judgmental probabilities of the conditions vary. To see how that goes, let's reformulate (4) in general terms:

> STAYING HOME. $pr(H|X = a) = f(a)$ if a is not in the interior of an interval that surely contains no values of X: ———$\cdots a \cdots$———

Here we must think of pr as a variable whose domain is a set of probability assignments that may assign different values to the condition $X = a$, but, for each H and each a that is not inside an excluded interval, assign the same value, $f(a)$, to the conditional probability of H given a.

[6]Carl G. Hempel, *Foundations of Natural Science,* Prentice-Hall (1966), p. 4.

Chapter 3

Testing Scientific Theories

Huygens gave this account of the scientific method in the introduction to his *Treatise on Light* (1690):

"... whereas the geometers prove their propositions by fixed and incontestable principles, here the principles are verified by the conclusions to be drawn from them; the nature of these things not allowing of this being done otherwise. It is always possible thereby to attain a degree of probability which very often is scarcely less than complete proof. To wit, when things which have been demonstrated by the principles that have been assumed correspond perfectly to the phenomena which experiment has brought under observation; especially when there are a great number of them, and further, principally, when one can imagine and foresee new phenomena which ought to follow from the hypotheses which one employs, and when one finds that therein the fact corresponds to our prevision. But if all these proofs of probability are met with in that which I propose to discuss, as it seems to me they are, this ought to be a very strong confirmation of the success of my inquiry; and it must be ill if the facts are not pretty much as I represent them."

In this chapter we interpret and extend Huygens's methodology in the light of the discussion of invariance, conditioning, and generalized conditioning in sections **2.1** and **2.2**.

3.1 Quantifications of Confirmation

The thought is that you see an episode of observation, experiment, or reasoning as confirming or infirming a hypotheses to a degree that depends on whether your probability for it increases or decreases during the episode, i.e.,

depending on whether your posterior probability, $new(H)$, is greater or less than your prior probability, $old(H)$. What Carnap[1] calls your DEGREE OF CONFIRMATION, $new(H) - old(H)$, is one measure of that change—positive for confirmation, negative for infirmation. Other measures are the *probability factor* and the *odds factor,* in both of which the turning point between confirmation and infirmation is 1 instead of 0. These are the factors by which old probabilities or odds are multiplied to get the new probabilities or odds.

PROBABILITY FACTOR:

$$\pi(H) =_{df} \frac{new(H)}{old(H)}$$

By your odds on one hypothesis *against* another—say, on G against an alternative H, is meant the ratio $pr(G)/pr(H)$ of your probability of G to your probability of H; by your odds simply "on G" is meant your odds on G against $\bar{G}$:

ODDS ON G AGAINST H $= pr(G)/pr(H)$; ODDS ON G $= pr(G)/pr(\neg G)$.

Your "odds factor"—or, better, "Bayes factor"—is the factor by which your old odds can be multiplied to get your new odds:

BAYES FACTOR (ODDS FACTOR):

$$\beta(G : H) = \frac{new(G)/new(H)}{old(G)/old(H)} = \frac{\pi(G)}{\pi(H)}$$

Where updating $old \mapsto new$ is by conditioning on D, your Bayes factor = your likelihood ratio:

If $new(\ \ ) = old(\ \ |D)$ then $\beta(G : H) = \dfrac{old(D|G)}{old(D|H)}$ = the LIKELIHOOD RATIO

A useful variant of the Bayes factor is its logarithm, dubbed by I. J. Good[2]

WEIGHT OF EVIDENCE:

$$w(G : H) =_{df} log\, \beta(G : H)$$

As odds vary from 0 to ∞, their logarithms ("lods") vary from $-\infty$ to $+\infty$. High probabilities (say, 100/101 and 1000/1001) correspond to widely spaced odds (100/1 and 1000/1); but low probabilities (say, 1/100 and 1/1000) correspond to odds (1/99 and 1/999) that are roughly as cramped as the probabilities. Going from odds to lods moderates the spread at the high end, and

[1]Rudolf Carnap, *Logical Foundations of Probability,* vol. 1; Chicago, 1950.

[2]I. J. Good, it Probability and the Weighing of Evidence, London, 1950, chapter 6.

treats high and low symmetrically. Thus, high odds 100:1 and 1000:1 become lods 2 and 3, and low odds 1:100 and 1:1000 become lods -2 and -3.

A. N. Turing was a playful, enthusiastic advocate of log odds, as his wartime code-breaking colleague I. J. Good reports:

"Turing suggested ... that it would be convenient to take over from acoustics and electrical engineering the notation of bels and decibels (db). In acoustics, for example, the bel is the logarithm to base 10 of the ratio of two intensities of sound. Similarly, if f [our β] is the factor in favour of a hypothesis, i.e., the ratio of its final to its initial odds, then we say that the hypothesis has gained $\log_{10} f$ bels or $10 \log_{10} f$ db. This may be described as the weight of evidence ... and $10 \log o$ db may be called the plausibility corresponding to odds o. Thus ... Plausibility gained = weight of evidence"[3]

"The weight of evidence can be added to the initial log-odds of the hypothesis to obtain the final log-odds. If the base of the logarithms is 10, the unit of weight of evidence was called a ban by Turing (1941) who also called one tenth of a ban a deciban (abbreviated to db). I hope that one day judges, detectives, doctors and other earth-ones will routinely weigh evidence in terms of decibans because I believe the deciban is an intelligence-amplifier."[4]

The reason for the name ban was that tens of thousands of sheets were printed in the town of Banbury on which weights of evidence were entered in decibans for carrying out an important classified process called Banburismus. A deciban or half-deciban is about the smallest change in weight of evidence that is directly perceptible to human intuition. ... The main application of the deciban was ... for discriminating between hypotheses, just as in clinical trials or in medical diagnosis. [Good (1979) p. 394][5]

Playfulness $\neq$ frivolity. Good is making a serious proposal, for which he claims support by extensive testing in the early 1940's at Bletchley Park, where the "Banburismus" language for hypothesis-testing was used in breaking each day's German naval "Enigma" code during World War II.[6]

---

[3] I. J. Good, op. cit., p. 63.

[4] I. J. Good, *Good Thinking* (Minneapolis, 1983), p. 132. Good means 'inteligence' in the sense of 'information': Think MI (military intelligence), not IQ.

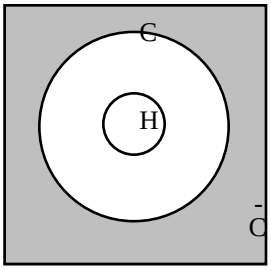
[5] I. J. Good, "A. M. Turing's Statistical Work in World War II" (*Biometrika* **66** (1979, 393-6), p. 394.

[6] See Andrew Hodges, *Alan Turing, the Enigma.* (New York, 1983). The dead hand of the British Official Secrets Act is loosening; see the A. N. Turing home page, http://www.turing.org.uk/turing/, maintained by Andrew Hodges.

3.2 Observation and Sufficiency

It only remains to recall the rationale for updating *old* to *new* by conditioning on C or on $\neg C$, i.e., setting $new(H) = old(H|\pm C)$ depending on whether what you observe assures you of C's truth or of its falsity. According to the analysis in sec. **3.1**, the warrant for this must be **sufficiency** for you of truth or falsity of C as evidence about H, which is to say, *rigidity,* $new(H|\pm C) = old(H|\pm C)$, so that that whatever information you acquire in the course of learnng that C is true over and above a bare report of C's truth or falsity has no further relevance to H. This is guaranteed if you have arranged for a statistician's stooge (sec. **2.1**) to send you the information about C in a one word telegram: 'true', or 'false'. But if the observers are the very people whose judgmental states are to be updated by the transition from *old* to *new*, the possibilty must be considered that the information about H conveyed by the observation will overflow the package provided by the sentence conveying the bare information that C is true, or that C is false.

EXAMPLE 1, **Huygens on light.**
Let H represent Huygens's hypotheses about light, and let C represent one of the 'new phenomena which ought to follow from the hypotheses which one employs'.



C (unshaded) is known to follow from *H*.

If we know that C follows from H, and we can discover by observation whether C is true or false, then we have the means to test H—more or less conclusively, depending on whether we find that C is false or true. If C proves false (shaded region), H is refuted decisively, for $H \wedge \neg C = \bot$. On the other hand, if C proves true, H's probability changes from

$$old(H) = \frac{\text{area of the H circle}}{\text{area of the } \top \text{ square}} = \frac{old(H)}{1}$$

to

$$new(H) = old(H|C) = \frac{\text{area of the H circle}}{\text{area of the C circle}} = \frac{old(H)}{old(C)}.$$

Then verification of C increases H's old probability by a factor $\pi(H) = 1/old(C)$; it is the antecedently least probable conclusions whose unexpected verification raises H's probability the most.[7] As Pólya put it: 'More danger, more honor.'[8]

In fact there can be no overflow if C is found to be false, for since the $\bar{C}$ region is disjoint from the H circle, *any* conditional probability function must assign 0 to H given falsity of C. This guarantees rigidity relative to $\bar{C}$: $new(H|\bar{C}) = old(H|\bar{C}) = 0$. No matter what else the observation might reveal about the circumstances of C's falsity, H would remain refuted.

It is in case of a positive result, verification of C, that overflow is possible. In this case, observation may provide further information that complicates matters by removing our warrant to update by conditioning on C.

EXAMPLE 2, **The Green Bean, yet again.**

H : This bean is lime-flavored.
C : This bean is green.

You are drawing a bean from a bag in which you know that half of the beans are green, all the lime-flavored ones are green, and the green ones are equally divided between lime and mint flavors. So before looking at the bean or tasting it, your probabilities are these:

$$old(C) = 1/2 = old(H|C);\ old(H) = 1/4.$$

But although $new(C) = 1$, your probability $new(H)$ for lime can drop below $old(H) = 1/4$ instead of rising to $1/2 = old(H|C)$ in case you yourself are the observer—for instance, if, when you see that the bean is green, you also get a whiff of mint, or also see that it has a special shade of green that you have found to be associated with the mint-flavored ones.

3.3 Leverrier on Neptune

We now turn to a methodological story more recent than Huygens's:[9]

"On the basis of Newton's theory, the astronomers tried to compute the motions of [...] the planet Uranus; the differences between theory and observation seemed to exceed the admissible limits of error. Some astronomers

[7]The corollary to Bayes's theorem in section **1.6** generalizes this result to cases where $pr(C|H) < 1$.

[8]George Pólya, Patterns of Plausible Inference, 2nd ed., Princeton University Press 1968, vol. 2, p. 126.

[9]George Pólya, op. cit., pp. 130-132.

suspected that these deviations might be due to the attraction of a planet revolving beyond Uranus' orbit, and the French astronomer Leverrier investigated this conjecture more thoroughly than his colleagues. Examining the various explanations proposed, he found that there was just one that could account for the observed irregularities in Uranus's motion: the existence of an extra-Uranian planet [sc., Neptune]. He tried to compute the orbit of such a hypothetical planet from the irregularities of Uranus. Finally Leverrier succeeded in assigning a definite position in the sky to the hypothetical planet [say, with a 1 degree margin of error]. He wrote about it to another astronomer whose observatory was the best equipped to examine that portion of the sky. The letter arrived on the 23rd of September 1846 and in the evening of the same day a new planet was found within one degree of the spot indicated by Leverrier. It was a large ultra-Uranian planet that had approximately the mass and orbit predicted by Leverrier."

We treated Huygens's conclusion as a strict deductive consequence of his principles. But Pólya made the more realistic assumption that Leverrier's prediction C (a bright spot near a certain point in the sky at a certain time) was highly probable but not 100%, given his H (namely, Newton's laws and observational data about Uranus). So $old(C|H) \approx 1$; and presumably the rigidity condition was satisfied so that $new(C|H) \approx 1$, too. Then verification of C would have raised H's probability by a factor $\approx 1/old(C)$, which is large if the prior probability $old(C)$ of Leverrier's prediction was ≈ 0.

Pólya offers a reason for regarding $1/old(C)$ as at least 180, and perhaps as much as 13131: The accuracy of Leverrier's prediction proved to be better than 1 degree, and the probability of a randomly selected point on a circle or on a sphere being closer than 1 degree to a previously specified point is 1/180 for a circle, and about 1/13131 for a sphere. Favoring the circle is the fact that the orbits of all known planets lie in the same plane ("the ecliptic"). Then the great circle cut out by that plane gets the lion's share of probability. Thus, if $old(C)$ is half of 1%, H's probability factor will be about 200.

3.4 Multiple Uncertainties

In Pólya's story, Leverrier loosened Huygens's tight hypothetico-deductive reasoning by backing off from deductive certainty to values of $old(C|H)$ falling somewhat short of 1—treating those values as approximately 1. Then the question arises: What are the effects of different shortfalls?

Similarly, we can back off from observational certainty to $new(C)$ values less

than 1. What if the confirming observation had raised the probability of Leverrier's C from a prior value of half of 1% to some posterior value short of 1?—say, $new(C) = 95\%$. Surely that would have increased H's probability by a factor smaller than Pólya's 200; but how much smaller?

And it would be more realistic to tell the story in terms of a point prediction with stated imprecision—say, ± 1 degree. (In fact, the new planet was observed within that margin: 57 seconds of arc from the point.) As between two theories that make such predictions, the one making the more precise prediction can be expected to gain the more from a confirming observation. But how much more? The following formula (due to John Burgess) for H's probability factor π when conditioning on C answers such questions, provided C and $\bar{C}$ satisfy the rigidity condition:

$$\pi(H, C) = 1 + \frac{new(C) - old(C) \times old(C|H) - old(C)}{old(C)old(\bar{C})}$$

By dint of lots of algebra, you can derive that formula from basic laws of probability and generalized conditioning (sec. **1.9**). If we call the term added to 1 in $\pi(H, C)$ the **strength of confirmation**, $sc(H, C)$, for H in view of C's change in probability, then we have

$$sc(H, C) = \pi(H, C) - 1$$

The sign ($\pm$) of $sc(H, C)$ distinguishes confirmation ($+$) from infirmation ($-$, "negative confirmation").

EXERCISES. What does $sc(H, C)$ reduce to in these cases?
(a) $new(C) = 1$
(b) $old(C|H) = 1$
(c) $new(C) = old(C|H) = 1$
(d) $old(C) = 0$ or 1, prior certainty about C.

To see the effect of precision, suppose that C predicts that a planet will be found within $\pm e$ of a certain point in the sky—a prediction that is definitely confirmed, within the limits of observational error. Thus $old(C|H) = new(C) = 1$, and $old(C)$ increases with e. Here $sc(H, C) = old(\bar{C})/old(C) =$ the prior odds against C, and H's probability factor is $1/old(C)$. Then if it was thought certain that the observed position would be in the plane of the ecliptic, $old(C)$ might well be proportional to e, $old(C) = ke$.
(e) On this assumption of proportionality, what happens to H's probability factor when E doubles?

3.5 Dorling on the Duhem Problem

Skeptical conclusions about the possibility of scientific hypothesis-testing are often drawn from the presumed arbitrariness of answers to the question of which to give up—theory, or auxiliary hypothesis—when they jointly contradict empirical data. The problem, addressed by Duhem in the first years of the 20th century, was agitated by Quine in mid-century. As drawn by some of Quine's readers, the conclusion depends on his assumption that aside from our genetical and cultural heritage, deductive logic is all we've got to go on when it comes to theory testing. That would leave things pretty much as Descartes saw them, just before the mid-17th century emergence in the hands of Fermat, Pascal, Huygens and others of the probabilistic ("Bayesian") methodology that Jon Dorling has brought to bear on various episodes in the history of science.

The conclusion is one that scientists themselves generally dismiss, thinking they have good reason to evaluate the effects of evidence as they do, but regarding formulation and justification of such reasons as someone else's job—the methodologist's. Here is an introduction to Dorling's work on that job, using extracts from his important but still unpublished 1982 paper[10]—which is reproduced in the web page "`http:\\www.princeton.edu\~bayesway`" with Dorling's permission.

The material is presented here in terms of odds factors. Assuming rigidity relative to D, the odds factor for a theory T against an alternative theory S that is due to learning that D is true will be the left-hand side of the following equation, the right-hand side of which is called "the likelihood ratio":

$$\text{Bayes Factor} = \text{Likelihood Ratio:}$$
$$\frac{old(T|D)/old(S|D)}{old(T)/old(S)} = \frac{old(D|T)}{old(D|S)}$$

The empirical result D is not generally deducible or refutable by T alone, or by S alone, but in interesting cases of scientific hypothesis testing D is

[10]Jon Dorling, 'Bayesian personalism, the methodology of research programmes, and Duhem's problem', *Studies in History and Philosophy of Science* **10**(1979)177-187. More along the same lines: Michael Redhead, 'A Bayesian reconstruction of the methodology of scientific research programmes,' *Studies in History and Philosophy of Science* **11**(1980)341-347. Dorling's unpublished paper from which excerpts appear here in sec. 2.7 - 2.10 is 'Further illustrations of the Bayesian solution of Duhem's problem' (29 pp., photocopied, 1982). References here ('sec. 4', etc.) are to the numbered sections of that paper. Dorling's work is also discussed in Colin Howson and Peter Urbach, *Scientific Reasoning: the Bayesian approach* (Open Court, La Salle, Illinois, 2nd ed., 1993).

deducible or refutable on the basis of the theory and an auxiliary hypothesis A (e.g., the hypothesis that the equipment is in good working order). To simplify the analysis, Dorling makes an assumption that can generally be justified by appropriate formulation of the auxiliary hypothesis:

Prior Independence

$$old(A \wedge T) = old(A)old(T),$$
$$old(A \wedge S) = old(A)old(S).$$

Generally speaking, S is not simply the denial of T, but a definite scientific theory in its own right, or a disjunction of such theories, all of which agree on the phenomenon of interest, so that, as an explanation of that phenomenon, S is a rival to T. In any case Dorling uses the independence assumption to expand the right-hand side of the Bayes Factor = Likelihood Ratio equation:

$$\beta(T : S) = \frac{old(D|T \wedge A)old(A) + old(D|T \wedge \neg A)old(\neg A)}{old(D|S \wedge A)old(A) + old(D|S \wedge \neg A)old(\neg A)}$$

To study the effect of D on A, he also expands $\beta(A : \bar{A})$ with respect to T (and similarly with respect to S, although we do not show that here):

$$\beta(A : \neg A) = \frac{old(D|A \wedge T)old(T) + old(D|A\neg T)old(\neg T)}{old(D|\neg A \wedge T)old(T) + old(D|\neg A \wedge \neg T)old(\neg T)}$$

3.5.1 Einstein/Newton, 1919

In these terms Dorling analyzes two famous tests that were duplicated, with apparatus differing in seemingly unimportant ways, with conflicting results: one of the duplicates confirmed T against S, the other confirmed S against T. But in each case the scientific experts took the experiments to clearly confirm one of the rivals against the other. Dorling explains why the experts were right:

'In the solar eclipse experiments of 1919, the telescopic observations were made in two locations, but only in one location was the weather good enough to obtain easily interpretable results. Here, at Sobral, there were two telescopes: one, the one we hear about, confirmed Einstein; the other, in fact the slightly larger one, confirmed Newton. Conclusion: Einstein was vindicated, and the results with the larger telescope were rejected.' (Dorling, sec. 4)

Notation
T: Einstein: light-bending effect of the sun
S: Newton: no light-bending effect of the sun
A: Both telescopes are working correctly
D: The conflicting data from both telescopes

In the odds factor $\beta(T : S)$ above, $old(D|T \wedge A) = old(D|S \wedge A) = 0$ since if both telescopes were working correctly they would not have given contradictory results. Then the first terms of the sums in numerator and denominator vanish, so that the factors $old(\neg T)$ cancel, and we have

$$\beta(T : S) = \frac{old(D|T \wedge \neg A)}{old(D|S \wedge \neg A)}$$

'Now the experimenters argued that one way in which A might easily be false was if the mirror of one or the other of the telescopes had distorted in the heat, and this was much more likely to have happened with the larger mirror belonging to the telescope which confirmed S than with the smaller mirror belonging to the telescope which confirmed T. Now the effect of mirror distortion of the kind envisaged would be to shift the recorded images of the stars from the positions predicted by T to or beyond those predicted by S. Hence $old(D|T \wedge \neg A)$ was regarded as having an appreciable value, while, since it was very hard to think of any similar effect which could have shifted the positions of the stars in the other telescope from those predicted by S to those predicted by T, $old(D|S \wedge \neg A)$ was regarded as negligibly small, hence the result as overall a decisive confirmation of T and refutation of S.' Thus the Bayes factor $\beta(T : S)$ is very much greater than 1.

3.5.2 Bell's Inequalities: Holt/Clauser

'Holt's experiments were conducted first and confirmed the predictions of the local hidden variable theories and refuted those of the quantum theory. Clauser examined Holt's apparatus and could find nothing wrong with it, and obtained the same results as Holt with Holt's apparatus. Holt refrained from publishing his results, but Clauser published his, and they were rightly taken as excellent evidence for the quantum theory and against hidden-variable theories.' (Dorling, sec. 4)

Notation
T: Quantum theory
S: Disjunction of local hidden variable theories

A: Holt's setup is sensitive enough to distinguish T from S
D: The specific correlations predicted by T and contradicted by S are not detected by Holt's setup

The characterization of D yields the first two of the following equations. In conjunction with the characterization of A it also yields $old(D|T \wedge \neg A) = 1$, for if A is false, Holt's apparatus was not sensitive enough to detect the correlations that would have been present according to T; and it yields $old(D|S \wedge \neg A) = 1$ because of the wild improbability of the apparatus "hallucinating" those specific correlations.[11]

$$old(D|T \wedge A) = 0,$$

$$old(D|S \wedge A) = old(D|T \wedge \neg A) = old(D|S \wedge \neg A) = 1$$

Substituting these values, we have

$$\beta(T, S) = old(\neg A)$$

Then with a prior probability of 4/5 for adequate sensitivity of Holt's apparatus, the odds between quantum theory and the local hidden variable theories shift strongly in favor of the latter, e.g., with prior odds 45:55 between T and S, the posterior odds are only 9:55, a 14% probability for T.

Now why didn't Holt publish his result?

Because the experimental result undermined confidence in his apparatus. Setting $\neg T = S$ in (2) because T and S were the only theories given any credence as explanations of the results, and making the same substitutions as in (4), we have

$$\beta(A : \neg A) = old(S)$$

so the odds on A fall from 4:1 to 2.2:1; the probability of A falls from 80% to 69%. Holt is not prepared to publish with better than a 30% chance that his apparatus could have missed actual quantum mechanical correlations; *the swing to S depends too much on a prior confidence in the experimental setup that is undermined by the same thing that caused the swing.*

Why did Clauser publish?

[11] Recall Venn on the credibilty of extraordinary stories: Problem 7 in sec. **1.11**.

Notation
T: Quantum theory
S: Disjunction of local hidden variable theories
C: Clauser's setup is sensitive enough
E: The specific correlations predicted by T and contradicted by S are detected by Clauser's setup

Suppose that $old(C) = .5$. At this point, although $old(A)$ has fallen by 11%, both experimenters still trust Holt's well-tried set-up better than Clauser's. Suppose Clauser's initial results E indicate presence of the quantum mechanical correlations pretty strongly, but still with a 1% chance of error. Then E strongly favors T over R:[12]

$$\beta(T,S) = \frac{old(E|T \wedge C)old(C) + old(E|T \wedge \neg C)old(\neg C)}{old(E|S \wedge C)old(C) + old(E|S \wedge \neg C)old(\neg C)} = 50.5$$

Starting from the low 9:55 to which T's odds fell after Holt's experiment, odds after Clauser's experiment will be 909:110, an 89% probability for T. The result E boosts confidence in Clauser's apparatus by a factor of

$$\beta(C : \neg C) = \frac{old(E|C \wedge T)old(T) + old(E|C \wedge S)old(S)}{old(E|\neg C \wedge T)old(T) + old(E|\neg C \wedge S)old(S)} = 15$$

This raises the initially even odds on C to 15:1, raises the probability from 50% to 94%, and lowers the 50% probability of the effect's being due to chance down to 6 or 7 percent.

3.5.3 Laplace/Adams

Finally, note one more class of cases: a theory T remains highly probable although (with auxiliary hypothesis A) it is incompatible with subsequently observed data D. With $S = \neg T$ in formulas (1) and (2), with $old(D|T \wedge A) = 0$ (so that the first terms in the numerators vanish), and setting

$$t = \frac{old(D|T\neg A)}{old(D|\neg T \wedge \neg A)}, \qquad s = \frac{old(D|\neg T \wedge A)}{old(D|\neg T \wedge \neg A)},$$

it is straightforward to show that

$$\beta(T : \neg T) = t/(s \times \frac{old(A)}{old(\neg A)} + 1),$$

[12]Numerically: $\frac{1\times.5+.01\times.5}{.01\times.5+.01\times.5} = 50.5$.

$$\beta(A : \neg A) = s/(t \times \frac{old(T)}{old(\neg T)} + 1),$$

$$\beta(T : A) = \frac{t}{s} \times \frac{old(\neg A)}{old(\neg T)}.$$

These formulas apply to 'a famous episode from the history of astronomy which clearly illustrated striking asymmetries in 'normal' scientists' reactions to confirmation and refutation. This particular historical case furnished an almost perfect controlled experiment from a philosophical point of view, because owing to a mathematical error of Laplace, later corrected by Adams, the same observational data were first seen by scientists as confirmatory and later as disconfirmatory of the orthodox theory. Yet their reactions were strikingly asymmetric: what was initially seen as a great triumph and of striking evidential weight in favour of the Newtonian theory, was later, when it had to be re-analyzed as disconfirmatory after the discovery of Laplace's mathematical oversight, viewed merely as a minor embarrassment and of negligible evidential weight against the Newtonian theory. Scientists reacted in the 'refutation' situation by making a hidden auxiliary hypothesis, which had previously been considered plausible, bear the brunt of the refutation, or, if you like, by introducing that hypothesis's negation as an apparently ad hoc face-saving auxiliary hypothesis.' (Dorling, sec. 1)

Notation
T: The theory, Newtonian celestial mechanics
A: The hypothesis that disturbances (tidal friction, etc.) make a negligible contribution to
D: The observed secular acceleration of the moon

Dorling argues on scientific and historical grounds for approximate numerical values

$$t = 1, \qquad s = \frac{1}{50}$$

The thought is that $t = 1$ because with A false, truth or falsity of T is irrelevant to D, and $t = 50s$ because in plausible partitions of $\neg T$ into rival theories predicting lunar accelerations, $old(R|\neg T) = 1/50$ where R is the disjunction of rivals not embarrassed by D. Then for a theorist whose odds are 3:2 on A and 9:1 on T (probabilities 60% for A and 90% for T),

$$\beta(T : \neg T) \quad = \quad \frac{100}{103}$$

$$\begin{aligned} \beta(A : \neg A) &= \frac{1}{500} \\ \beta(T : A) &= 200 \end{aligned}$$

Thus the prior odds 900:100 on T barely decrease, to 900:103; the new probability of T, 900/1003, agrees with the original 90% to two decimal places. But odds on the auxiliary hypothesis A drop sharply, from prior 3:2 to posterior 3/1000, i.e., the probability of A drops from 60% to about three tenths of 1%; odds on the theory against the auxiliary hypothesis increase by a factor of 200, from a prior value of 3:2 to a posterior value of 300:1.

3.5.4 Dorling's Conclusions

'Until recently there was no adequate theory available of how scientists should change their beliefs in the light of evidence. Standard logic is obviously inadequate to solve this problem unless supplemented by an account of the logical relations between degrees of belief which fall short of certainty. Subjective probability theory provides such an account and is the simplest such account that we posess. When applied to the celebrated Duhem (or Duhem-Quine) problem and to the related problem of the use of ad hoc, or supposedly ad hoc, hypotheses in science, it yields an elegant solution. This solution has all the properties which scientists and philosophers might hope for. It provides standards for the validity of informal deductive reasoning comparable to those which traditional logic has provided for the validity of informal deductive reasoning. These standards can be provided with a rationale and justification quite independent of any appeal to the actual practice of scientists, or to the past success of such practices.[13] Nevertheless they seem fully in line with the intuitions of scientists in simple cases and with the intuitions of the most experienced and most successful scientists in trickier and more complex cases. The Bayesian analysis indeed vindicates the rationality of experienced scientists' reactions in many cases where those reactions were superficially paradoxical and where the relevant scientists themselves must have puzzled over the correctness of their own intuitive reactions to the evidence. It is clear that in many such complex situations many less experienced commentators and critics have sometimes drawn incorrect conclusions

[13]Here a long footnote explains the Putnam-Lewis Dutch book argument for conditioning. Putnam stated the result, or, anyway, a special case, in a 1963 *Voice of America* Broadcast, 'Probability and Confirmation', reprinted in his *Mathematics, Matter and Method,* Cambridge University Press (1975) 293-304. Paul Teller, 'Conditionalization and observation', *Synthese* **26** (1973) 218-258, reports a general argument to that effect which Lewis took to have been what Putnam had in mind.

and have mistakenly attributed the correct conclusions of the experts to scientific dogmatism. Recent philosophical and sociological commentators have sometimes generalized this mistaken reaction into a full-scale attack on the rationality of men of science, and as a result have mistakenly looked for purely sociological explanations for many changes in scientists' beliefs, or the absence of such changes, which were in fact, as we now see, rationally de rigeur.' (Dorling, sec. 5)

3.6 Old News Explained

In sec. **3.2** we analyzed the case Huygens identified as the principal one in his *Treatise on Light:* A prediction C long known to follow from a hypothesis H is now found to be true. Here, if the rigidity condition is satisfied, $new(H) = old(H|C)$, so that the probability factor is $\pi(H) = 1/old(C)$.

But what if some long-familiar phenomenon C, a phenomenon for which $old(C) = 1$, is newly found to follow from H in conjunction with familiar background information B about the solar system, and thus to be explained by $H \wedge B$? Here, if we were to update by conditioning on C, the probability factor would be 1 and $new(H)$ would be the same as $old(H)$. No confirmation here.[14]

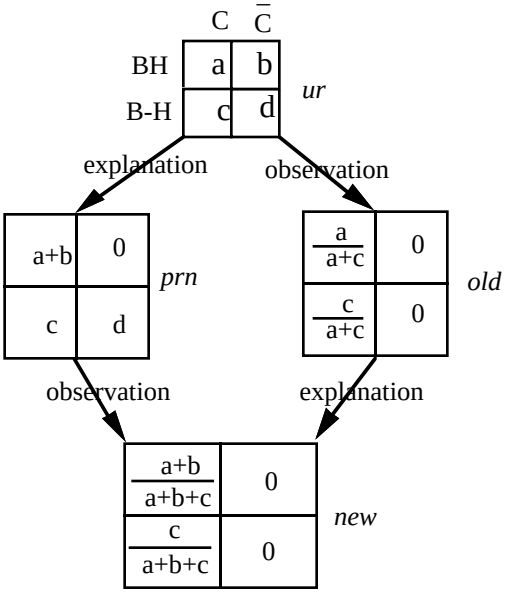
Wrong, says Daniel Garber:[15] The information prompting the update is not that C is true, but that $H \wedge B$ implies C. To condition on that news Garber proposes to to enlarge the domain of the probability function *old* by adding to it the hypothesis that C follows from $H \wedge B$ together with all further truth-functional compounds of that new hypothesis with the *old* domain. Using some extension old^* of *old* to the enlarged domain, we might then have $new(H) = old^*(H|H \wedge B$ implies $C)$. That is an attractive approach to the problem, if it can be made to work.[16]

The alternative approach that we now illustrate defines the *new* assignment on the same domain that *old* had. It analyzes the $old \mapsto new$ transition by embedding it in a larger process, $ur \overset{obs}{\mapsto} old \overset{expl}{\mapsto} new$, in which *ur* represents an original state of ignorance of C's truth and its logical relationship to H:

[14]This is what Clark Glymour has dubbed 'the paradox of old evidence': see his *Theory and Evidence,* Princeton University Press, 1980.

[15]See his 'Old evidence and logical omniscience in Bayesian confirmation theory' in *Testing Scientific Theories,* ed. John Earman, University of Minnesota Press, 1983. For further discussion of this proposal see 'Bayesianism with a human face' in my *Probability and the Art of Judgment,* Cambridge University Press, 1992.

[16]For a critique, see chapter 5 of John Earman's *Bayes or Bust?* (MIT Press, 1992).



Confirmation by New Explanation

EXAMPLE 1, **The perihelion of Mercury.**

H: GTR applied to the Sun-Mercury system
C: Advance of 43 seconds of arc per century[17]

In 1915 Einstein presented a report to the Prussian Academy of Sciences explaining the then-known advance of approximately 43″ per century in the perihelion of Mercury in terms of his ("GTR") field equations for gravitation. An advance of 38″ per century had been reported by Leverrier in 1859, due to 'some unknown action on which no light has been thrown'.[18]

> 'The only way to explain the effect would be ([Leverrier] noted) to increase the mass of Venus by at least 10 percent, an inadmissible modification. He strongly doubted that an intramercurial planet ["Vulcan"], as yet unobserved, might be the cause. A swarm of intramercurial asteroids was not ruled out, he believed.'

Leverrier's figure of 38″ was soon corrected to is present value of 43″, but the difficulty for Newtonian explanations of planetary astronomy was still

[17] Applied to various Sun-planet systems, the GTR says that all planets' perihelions advance, but that Venus is the only one for which that advance should be observable. This figure for Mercury has remained good since its publication in 1882 by Simon Newcomb.

[18] *'Subtle is the Lord...', the Science and Life of Albert Einstein,* Abraham Pais, Oxford University Press, 1982, p. 254.

in place 65 years later, when Einstein finally managed to provide a general relativistic explanation 'without special assumptions' (such as Vulcan)—an explanation which was soon accepted as strong confirmation for the GTR.

In the diagram above, the left-hand path, $ur \mapsto prn \mapsto new$, represents an expansion of the account in sec. **3.2** of Huygens's "principal" case ('*prn*' for 'principal'), in which the confirming phenomenon is verified *after* having been predicted via the hypothesis which its truth confirms. The "*ur*" probability distribution, indicated schematically at the top, represents a time before C was known to follow from H. To accomodate that discovery the 'b' in the *ur*-distribution is moved left and added to the 'a' to obtain the top row of the *prn* distribution. The reasoning has two steps. First: Since we now know that H is true, C must be true as well. Therefore $prn(H \wedge \neg C)$ is set equal to 0. Second: Since implying a prediction that may well be false neither confirms nor infirms H, the *prn* probability of H is to remain at its *ur*-value even though the probability of $H \wedge \neg C$ has been nullified. Therefore the probability of $H \wedge C$ is increased to $prn(H \wedge C) = a + b$. And this *prn* distribution is where the Huygens on light example in sec. **2.2** *begins,* leading us to the bottom distribution, which assigns the following odds to H:

$$\frac{new(H)}{new(\neg H)} = \frac{a+b}{c}.$$

In terms of the present example this *new* distribution is what Einstein arrived at from the distribution labelled *old*, in which

$$\frac{old(H)}{old(\neg H)} = \frac{a}{c}.$$

The rationale is

COMMUTATIVITY:
The *new* distribution is the same, no matter if the observation or the explanation comes first.

Now the Bayes factor gives the clearest view of the transition $old(H) \mapsto new(H)$:[19]

[19]The approach here is taken from Carl Wagner, "Old evidence and new explanation III", *PSA 2000* (J. A. Barrett and J. McK. Alexander, eds.), part 1, pp. S165-S175 (Proceedings of the 2000 biennial meeting of the Philosophy of Science Association, supplement to *Philosophy of Science* **68** [2001]), which reviews and extends earlier work, "Old evidence and new explanation I" and "Old evidence and new explanation II" in *Philosophy of Science* **64,** No. 3 (1997) 677-691 and **66** (1999) 283-288.

$$\beta(H : \neg H) = \frac{a+b}{c} / \frac{a}{c} = 1 + \frac{ur(\neg C|H)}{ur(C|H)}.$$

Thus the new explanation of old news C increases the odds on H by a factor of (1 + the *ur*-odds against C, given H). Arguably, this is very much greater than 1, since, in a notation in which $C = C_{43}$ = an advance of $(43 \pm .5)''/c$ and similarly for other C_i's, $\neg C$ is a disjunction of very many "almost" incompatible disjuncts: $\neg C = \ldots C_{38} \vee C_{39} \vee C_{40} \vee C_{41} \vee C_{42} \vee C_{44} \vee \ldots$.[20] That's the good news.[21]

And on the plausible assumption[22] of *ur*-independence of H from C, the formula for β becomes even simpler:

$$\beta(H : \neg H) = 1 + \frac{ur(\bar{C})}{ur(C)}$$

The bad news is that *ur* and *prn* are new constructs, projected into a mythical methodological past. But maybe that's not so bad. As we have just seen in a very sketchy way, there seem to be strong reasons for taking the ratio $ur(\neg C|H)/ur(C|H)$ to be very large. This is clearer in the case of *ur*-independence of H and C:

EXAMPLE 2, **Ur-independence.**
If Wagner's independence assumption is generally persuasive, the physics community's *ur*-odds against C ("C_{43}") will be very high, since, behind the veil of *ur*-ignorance, there is a large array of C_i's, which, to a first approximation, are equiprobable. Perhaps, 999 of them? Then $\beta(H : \neg H) \geq 1000$, and the weight of evidence (**2.1**) is $w(H : \neg H) \geq 3$.

3.7 Problems

1 "Someone is trying decide whether or not T is true. He notes that T is a consequence of H. Later he succeeds in proving that H is false. How does this refutation affect the probability of T?"[23]

2 "We are trying to decide whether or not T is true. We derive a sequence of consequences from T, say $C_1, C_2, C_3, \ldots$. We succeed in verifying C_1, then C_2,

[20] "Almost": $ur(C_i \wedge C_j) = 0$ for any distinct integers i and j.
[21] "More [*ur*]danger, more honor." See Pólya, quoted in sec. **2.2** here.
[22] Carl Wagner's, again.
[23] This problem and the next are from George Pólya, 'Heuristic reasoning and the theory of probability', *American Mathematical Monthly* **48** (1941) 450-465.

then C_3, and so on. What will be the effect of these successive verifications on the probablity of T?"

3 Fallacies of "yes/no" confirmation.[24] Each of the following plausible rules is unreliable. Find counterexamples—preferably, simple ones.
(a) If D confirms T, and T implies H, then D confirms H.
(b) If D confirms H and T separately, it must confirm their conjunction, TH.
(c) If D and E each confirm H, then their conjunction, DE, must also confirm H.
(d) If D confirms a conjunction, TH, then it cannot infirm each conjunct separately.

4 Hempel's ravens. It seems evident that black ravens confirm (H) 'All ravens are black' and that nonblack nonravens do not. Yet H is logically equivalent to 'All nonravens are nonblack'. Use probabilistic considerations to resolve this paradox of "yes/no" confirmation.[25]

6 Show that the boxed formulas in **2.8** hold.

7 A condition equivalent to rigidity
Show that if pr_1 and pr_2 are probability assignments to $\mathcal{A}$=$\{H \wedge C, H \wedge \neg C, \neg H \wedge C, \neg H \wedge \neg C\}$, then the update $pr_1 \mapsto pr_2$ satisfies (a) if and only if it satisfies (b):
(a) $pr_1(C| \pm H) = pr_2(C| \pm H)$
(b) $pr_1(A) = pr_2(A')$ for every $A, A' \in \mathcal{A}$

8 Wagner III (sec **3.6**).[26] Carl Wagner extends his treatment of old evidence newly explained to cases in which updating is by generalized conditioning on a countable sequence $\mathcal{C} = \langle C_1, C_2, \ldots \rangle$ where none of the updated probabilies need be 1, and to cases in which each of the $\mathcal{C}$'s "follows only probabilistically" from H in the sense that the conditional probabilities of the $\mathcal{C}$'s, given H, are high but < 1. Here we focus on the simplest case, in which $\mathcal{C}$ has two members, $\mathcal{C} = \langle C, \neg C \rangle$. Where (as in the diagram in sec. **3.6**) updates are not always from *old* to *new*, we indicate the prior and posterior probability measures by a subscript and superscript: β^{post}_{prior}. In particular,

[24]These stem from Carl G. Hempel's 'Studies in the logic of confirmation', *Mind* **54** (1945) 1-26 and 97-121, which is reprinted in his *Aspects of Scientific Explanation,* The Free Press, New York, 1965.

[25]The paradox was first floated (1937) by Carl G. Hempel, in an abstract form: See pp. 50-51 of his *Selected Philosophical Essays,* Cambridge University Press, 2000. For the first probabilistic solution, see "On confirmation" by Janina Hosiasson-Lindenbaum, *The Journal of Symbolic Logic* **5** (1940) 133-148. For a critical survey of more recent treatments, see pp. 69-73 of John Earman's *Bayes or Bust?*

[26]This is an easily detatchable bit of Wagner's "Old evidence and new explanation III" (cited above in sec. 3.6).

we shall write $\beta^{new}_{old}(H : \neg H)$ explicitly, instead of $\beta(H : \neg H)$ as in (3) and (4) of sec. **3.6**. And we shall use 'β^*' as shorthand: $\beta^* = \beta^{old}_{ur}(C : \neg C)$.

(a) Show that in the framework of sec. **3.6** COMMUTATIVITY is equivalent to UNIFORMITY:
$\beta^{new}_{old}(A : A') = \beta^{prn}_{ur}(A : A')$ if $A, A' \in \{H \wedge C,\ H \wedge \neg C,\ \neg H \wedge C,\ \neg H \wedge \neg C\}$.

(b) Show that uniformity holds whenever the updates $ur \mapsto old$ and $prn \mapsto new$ are both by generalized conditioning.

Now verify that where uniformity holds, so do the following two formulas:

(c) $\beta^{new}_{ur}(H : \neg H) = \dfrac{(\beta^* - 1)prn(C|H) + 1}{(\beta^* - 1)prn(C|\neg H) + 1}$

(d) $\beta^{old}_{ur}(H : \neg H) = \dfrac{(\beta^* - 1)ur(C|H) + 1}{(\beta^* - 1)ur(C|\neg H) + 1}$

And show that, given uniformity,

(e) if H and C are ur-independent, then $old(H) = ur(H)$ and, therefore,

(f) $\beta^{new}_{old}(H : \neg H) = \dfrac{(\beta^* - 1)prn(C|H) + 1}{(\beta^* - 1)prn(C|\neg H) + 1}$, so that

(g) Given uniformity, if H and C are ur-independent then

(1) $\beta^{new}_{old}(H : \neg H)$ depends only on β^*, $prn(C|H)$ and $prn(C|\neg H)$;

(2) if $\beta > 1$ and $prn(C|H) > prn(C|\neg H)$, then $\beta^{new}_{old}(H : \neg H) > 1$;

(3) $\beta^{new}_{old}(H : \neg H) \to \dfrac{prn(C|H)}{prn(C|\neg H)}$ as $\beta \to \infty$.

Chapter 4

Expectation Primer

For the 17th century founders of modern probability theory, "expectation" (as in 'life expectancy') was the basic concept, in terms of which probability was understood. These days, the order is usually reversed, expectation being defined in terms of probability; but in many ways, the old order is smoother:[1]

4.1 Probability and Expectation

Your **expectation** $E(X)$ of a magnitude (or **"random variable")** X is usually defined as a weighted average of the values you think it can assume, in which the weights are your probabilities for those values. In particular, if X can assume only finitely many values, x_i, its expecatation is the sum[2]

$$E(X) = \sum_i x_i \cdot pr(X = x_i) \qquad (pr \mapsto E)$$

If the magnitude has only only one possible value, the "random variable" is a constant. And if it can assume more than one value but for some i your $pr(X = x_i) = 1$, the magnitude is a constant from your point of view.)

Example 1, **Giant Pandas.** Let $X =$ the birth weight to the nearest pound of the next giant panda to be born in captivity. If you have definite probabilities $p_0, \ldots, p_{99}$ for the hypotheses that $X = 0, \ldots, 99$, your expectation of X will be $0 \cdot p_0 + \ldots + 99 \cdot p_{99}$.[3]

[1] For updated versions of it, see Bruno de Finetti's *Theory of Probability,* volume 1 (Wiley, 1975) and Peter Whittle's *Probability via Expectation* (3rd edition, Springer, 1992).

[2] If X is capable of taking on any real value in the interval from a to b then $E(X) = \int_a^b x\, dF(x)$, where $F(x) = pr(X < x)$.

[3] If you assign probability 0 to values of X greater than 99, the '$+\ldots$' can be dropped.

But we could equally well have defined your probabilities as your expectations of particularly simple random variables called "indicators". The ***indicator*** of a hypothesis H is a random variable, $\boldsymbol{I_H}$, which takes the value 1 or 0 depending on whether H is true or false:

$$I_H = \begin{cases} 1 \text{ if } H \text{ is true} \\ 0 \text{ if } H \text{ is false} \end{cases}$$

Then we can define

$$pr(H) = E(I_H). \qquad (E \mapsto pr)$$

Note that as 1 and 0 are the only values I_H can assume, the $pr \mapsto E$ definition 1 gives us $E(I_H) = 1 \cdot pr(H) + 0 \cdot pr(\neg H)$, which $= pr(H)$.

Observe that your expectation of a random variable need not be one of the values you think it can assume. Thus, in the Panda example, X must be a whole number; but its expected value, which is an average of whole numbers, need not be a whole number. Nor need your expectation of the indicator of past life on Mars be one of the values that indicators can assume, i.e., 0 or 1; it might well be $\frac{1}{10}$, as in the story at the beginning of chapter 1.

4.1.1 Exchangeability and Independence[4]

Exchangeability and independence of random variables are defined in terms of the corresponding properties (sec. 1.7 and 1.8) of propositions about their values. In the **"discrete"** case, where each X_i can take only a countable number of different values, it is sufficient to consider propositions assigning those values:

> Random variables $X_1, \ldots, X_n$, each of which takes only countably many different values, are said to be exchangeable (or independent) when the propositions $X_1 = x_1, \ldots, X_n = x_n$ are exchangeable (or independent) for all combinations of values x_i of the X_i.

More generally, where the possible values might be all of the real numbers, independence and exchangeability of the X_i are defined as independence and exchangeability of propositions of form $X_i \leq r_i$:

> Real-valued random variables $X_1, \ldots, X_n$ are said to be exchangeable (or independent) when, for every n-tuple $r_1, \ldots, r_n$ of real numbers, the n propositions of form $X_i \leq r_i$ are exchangeable (or independent).

[4]Bruno de Finetti, *Theory of Probability,* volume 2 (Wiley, 1974), pp. 161-2.

4.1.2 Independence and Noncorrelation[5]

Random variables X, Y are said to be **uncorrelated** from your point of view when your expectation of their product is simply the product of your separate expectations: $E(XY) = E(X)E(Y)$. In the special case where X and Y are the indicators of propositions, noncorrelation reduces to independence: $E(XY) = E(X)E(Y)$ iff $pr(X = x \wedge Y = y) = pr(X = x)pr(Y = y)$ for all four pairs (x, y) of zeroes and ones.[6] But in general the uncorrelated pairs of random variables are not simply the independent ones.

EXAMPLE 1, **Non-correlation without independence.**[7] You assign equal probabilities to the 27 ways there are of placing tickets a, b, c in three bins. If $Y = 1, 2, 3$ is the number of occupied bins and $X = 0, 1, 2, 3$ is the number of tickets in the first bin, your probabilities for the 12 pairs of values will be

	$X = 0$	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	2/27	0	0	1/27
$Y = 2$	6/17	6/27	6/27	0
$Y = 3$	0	6/27	0	0

Crunching three times through the $pr \mapsto E$ formula at the start of this section you will find that $E(X) = 1$, $E(Y) = \frac{19}{9}$ and $E(XY) = \frac{19}{9} = E(X)E(Y)$.

EXAMPLE 2, **Non-correlation in spite of deterministic dependency.**[8] You have probability 1/4 for each of X = -2, -1, 1, 2; and $Y = X^2$, so that you have probability 1/2 for each possible value of Y (= 1 or 4), and probability 1/4 for each conjunction $X = x \wedge Y = x^2$ with $x = -2, -1, 1, 2$. Then you see X and Y as uncorrelated in the technical sense, for $E(X) = 0$ and $E(XY) = \frac{-8-1+1+8}{4} = 0 = E(X)E(Y)$.

Then noncorrelation need not imply independence. But independence does always imply noncorrelation—e.g., if X and Y take only finitely many values:[9]

> If $pr(X = x \wedge Y = y) = pr(X = x) \cdot pr(Y = y)$ for all values x of X and y of Y, then $E(XY) = E(X)E(Y)$.

[5] William Feller, *Theory of Probability,* vol. 1 (Wiley, 2nd ed., 1957).

[6] A simpler, equivalent condition is $pr(X = 1 \wedge Y = 1) = pr(X = 1)pr(Y = 1)$.

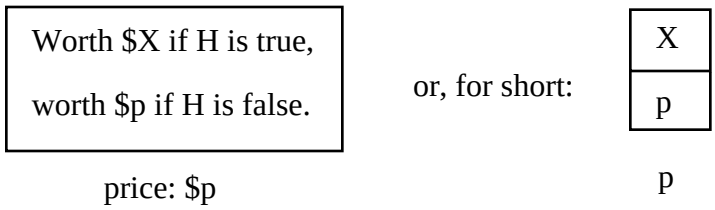
[7] Feller, p. 201, example (a).

[8] Feller, p. 222, example (a).

[9] *Proof.* By the $pr \mapsto E$ formula, $E(XY) = \sum_{i,j} x_i y_j \, pr(X = x_i \wedge Y = y_j)$, or, by independence, $\sum_{i,j} x_i y_j \, pr(X = x_i) \, pr(Y = y_j)$. By the same formula, $E(X)E(Y) = (\sum_i x_i \, pr(X = x_i))(\sum_j y_j \, pr(Y = y_j))$, which $= \sum_{i,j} x_i y_j \, pr(X = x_i) \, pr(Y = y_j)$ again.

4.2 Conditional Expectation

Just as we defined your conditional probabilities as your buying-or-selling prices for tickets that represent conditional bets, so we define your conditional expectation of a random variable X given truth of a statement H as your buying-or-selling price for the following ticket:



We write '$E(X|H)$' for your conditional expectation of X given H.

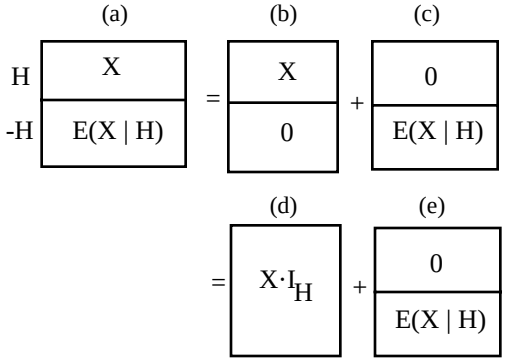
The following rule might be viewed as defining conditional expectations as quotients of unconditional ones when $pr(H) \neq 0$, just as the quotient rule for probabilities might be viewed as defining $pr(G|H)$ as $pr(GH)/pr(H)$.

$$\textbf{Quotient Rule:} \quad E(X|H) = \frac{E(X \cdot I_H)}{E(I_H)} = \frac{E(X \cdot I_H)}{P(H)} \text{ if } pr(H) \neq 0.$$

Of course this can be written as follows, without the condition $pr(H) \neq 0$.

$$\textbf{Product Rule:} \quad E(X \cdot I_H) = E(X|H) \cdot pr(H)$$

A "Dutch book" consistency argument can be given for this product rule that is simpler than the one given in sec. 1.4 for the probability product rule: Consider the following five tickets.



Clearly, ticket (a) has the same dollar value as (b) and (c) together on each hypothesis as to whether H is true, or false. And (d) and (e) have the same

values as (b) and (c), since $\mathrm{X}{\cdot}I_H = \begin{cases} X \text{ if } H \text{ is true} \\ 0 \text{ if } \neg H \text{ is true} \end{cases}$ and (e) = (c). Then unless your price for (a) is the sum of your prices for (d) and (e), so that the condition $E(X|H) = E(X \cdot I_H) + E(X|H)pr(\neg H)$ is met, you are inconsistently placing different values on the same prospect depending on whether it is described in one or the other of two provably equivalent ways. Now setting $P(\neg H) = 1 - P(H)$ and simplifying, the condition boils down to the product rule.

Historical Note. In the 18th century Thomas Bayes defined probability in terms of expectation as follows:

> "The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening."[10]

This is simply the product rule, solved for $P(H) = \dfrac{E(X \cdot I_H)}{E(X|H)}$. Explanation: $E(X \cdot I_H)$ is "an expectation [$\$X$] depending on the happening of the event". $\$E(X|H)$ is "the value of the thing expected upon its [H's] happening".

4.3 Laws of Expectation

The axioms for expectation are the product rule and

$$\textbf{Linearity: } E(aX + bY + c) = aE(X) + bE(Y) + c$$

Three notable special cases of linearity are obtained by setting a, b, c equal to $1, 1, 0$ (additivity), or setting $b = c = 0$ (proportionality) or $a = b = 0$ (constancy):

$$\begin{aligned} \textbf{Additivity}: &\quad E(X + Y) = E(X) + E(Y) \\ \textbf{Proportionality}: &\quad E(aX) = aE(X) \\ \textbf{Constancy}: &\quad E(c) = c \end{aligned}$$

If n hypotheses H_i are well-defined, they have actual truth values (truth or falsehood) even if we do not know what they are. Therefore we can define

[10] 'Essay toward solving a problem in the doctrine of chances,' *Philosophical Transactions of the Royal Society* 50 (1763), p. 376, reprinted in *Facsimiles of Two Papers by Bayes*, New York: Hafner, 1963.

the **number of truths** among them as $I_{H_1} + \ldots + I_{H_n}$. If you have definite expectations $E(I_{H_i})$ for them, you will have a definite expectation for that sum, which, by additivity, will be $E(I_{H_1}) + \ldots + E(I_{H_1})$.

EXAMPLE 2, **Calibration.** Suppose you attribute the same probability, p, to success on each trial of a certain experiment: $pr(H_1) = \ldots = pr(H_n) = p$. Consider the indicators of the hypotheses, H_i, that the different trials succeed. The number of successes in the n trials will be the unknown sum of the indicators, so the unknown **success rate** will be $\boldsymbol{S_n} = \frac{1}{n}\sum_{i=1}^{n} H_i$. Now by additivity and constancy, your expectation of the success rate must be p:

$$E(\frac{1}{n}\sum_{i=1}^{n} I_{H_i}) = \frac{1}{n}E(\sum_{i=1}^{n} I_{H_i}) = \frac{1}{n}E(\sum_{i=1}^{n} pr(H_i)) = \frac{1}{n}(np) = p.$$

The term 'calibration' comes from the jargon of weather forecasting; forecasters are said to be well calibrated (say, last year, for rain, at the level $p = .8$) when, last year, it rained on about 80% of the days for which the forecaster turns out to have said there would be an 80% chance of rain.[11]

Why Expectations are Additive. Suppose x and y are your expectations of magnitudes X and Y—say, rainfall in inches during the first and second halves of next year—and z is your expectation for next year's total rainfall, $X + Y$. Why should z be $x + y$? The answer is that in every eventuality about rainfall at the two locations, the value of the first two of these tickets together is the same as the value of the third:

$$\text{Values: } \boxed{\$X} + \boxed{\$Y} = \boxed{\$(X+Y)}$$
$$\text{Prices: } \$x \ + \ \$y \ = \ \$(x+y)$$

Then unless the prices you would pay for the first two add up to the price you would pay for the third, you are inconsistently placing different values on the same prospect, depending on whether it is described to you in one or the other of two provably equivalent ways.

The magnitudes of which we have expectations are often called **'random variables',** because they may have various values with various probabilities. *May* have: in the linearity property a, b, c are constants, so that, as we have seen, $E(c)$ makes sense, and equals c. But more typically, a random variable might have any of a number of values as far as you know. Convexity says that

[11] For more about calibration, etc., see Morris DeGroot and Stephen Fienberg, "Assessing Probability Assessors: Calibration and Refinement," in Shanti S. Gupta and James O. Berger (eds.), *Statistical Decision Theory and Related Topics III*, Vol. 1, New York: Academic Press, 1982, pp. 291-314.

your expectation for the variable cannot be larger than all of those values, or smaller than all of them:[12]

> **Convexity:** $E(X)$ lies in the range from the largest to the smallest values that X can assume.

The following connection between conditional and unconditional expectations is of particular importance. Here, the H's are any hypotheses whatever:

> **Law of Total Expectation:** If no two of $H_1, H_2, \ldots$ are compatible and H is their disjunction, then $E(X|H) = E(X|H_1)pr(H_1) + E(X|H_2)pr(H_2) + \ldots$.

Proof: $X \cdot I_H = X \cdot I_{H_1} + X \cdot I_{H_2} + \ldots$; apply E to both sides, then use additivity and the product rule.

When conditions are certainties for you, conditional expectations reduce to unconditional ones:

> **Certainty:** $E(X|H) = E(H)$ if $pr(H) = 1$.

Note that on a context of form $E(\cdots Y \cdots | Y =?)$ it is always permissible to rewrite Y as ? at the left. Example: $E(Y^2|Y = 2x) = E(4X^2|Y = 2x)$.

Applying conditions:

$$E(\cdots Y \cdots | Y =?) = E(\cdots ? \cdots | Y =?) \quad \textbf{(OK!)}$$

But we cannot generally discharge a condition $Y =?$ by rewriting Y as ? at the left *and dropping the condition.* Example: $E(Y^2 \,|\, Y = 2X)$ cannot be relied upon to equal$E(4X^2)$.

The Discharge Fallacy:

$$E(\cdots Y \cdots | Y =?) = E(\cdots ? \cdots) \quad \textbf{(NOT!)}$$

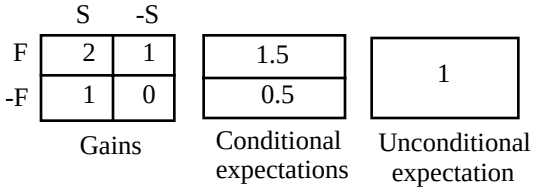
Example 3, **The problem of the two sealed envelopes.** One contains a check for an unknown whole number of dollars, the other a check for twice or half as much. Offered a free choice, you pick one at random. What is wrong with the following argument for thinking you should have chosen the other?

> "Let X and Y be the values of the checks in the one and the other. As you think Y equally likely to be $.5X$ or $2X$, $E(Y)$ will be $.5E(.5X) + .5E(2X) = 1.25E(X)$, which is larger than $E(X)$."

[12] Where X can assume only a finite numer of values, convexity follows from linearity.

4.4 Physical Analogies; Mean and Median

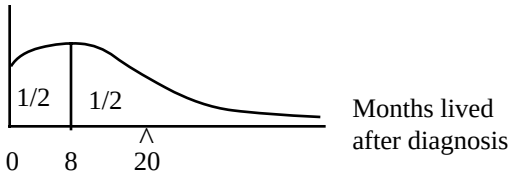
Hydraulic Analogy. Let 'F' and 'S' mean heads on the first and second tosses of an ordinary coin. Suppose you stand to gain a dollar for each head. Then your net gain in the four possibilities for truth and falsity of F and S will be as shown at the left below.



Think of that as a map of flooded walled fields in a plain, with the numbers indicating water depths in the four sections. In the four regions, depths are values of X and areas are probabilities. To find your expectation for X, open sluices so that the water reaches a common level in all sections. That level will be $E(X)$. To find your conditional expectation for X given F, open a sluice between the two sections of F so that the water reaches a single level in the two. That level will be $E(X|F)$, which is, 1.5 in the diagram. Similarly, $E(X|\neg F) = 0.5$. To find your unconditional expectation of gain, open all four sluices so that the water reaches the same level throughout: $E(X) = 1$.

But there is no mathematical reason for magnitudes X to have only a finite numbers of values, e.g., we might think of X as the birth weight in pounds of the next giant panda to be born in captivity—to no end of decimal places of accuracy, as if that meant something.[13] Nor is there any reason why X cannot have negative values, as in the case of temperature.

Balance. The following analogy is more easily adapted to the continuous case. On a weightless rigid beam, positions represent values of a magnitude X that might go negative as well as positive. Pick a zero, a unit, and a positive direction on the beam. Get a pound of modelling clay, and distribute it along the beam so that the weight of clay on each section represents your probability that the true value of X is in that section—say, like this:



[13]It does not. The commonplace distinction between panda and ambient moisture, dirt, etc. is not drawn finely enough to let us take the remote decimal places seriously.

"The Median is not the Message."[14] In 1985 Stephen Gould wrote:

> "In 1982, I learned I was suffering from a rare and serious cancer. After surgery, I asked my doctor what the best technical literature on the cancer was. She told me ... that there was nothing really worth reading. I soon realized why she had offered that humane advice: my cancer is incurable, with a median mortality of eight months after discovery."

Locations on the beam are months lived after diagnosis; the weight of clay on the interval from 0 to m is the probability of still being alive in m months.

In terms of the beam analogy here are the key definitions, of the terms "median" and "mean"—the latter being a synonym for "expectation":

> The **median** is the point on the beam that divides the weight of clay in half: the probabilities are equal that the true value of X is represented by a point to the right and to the left of the median.
>
> The **mean** (= your expectation) is the point of support at which the beam would just balance.

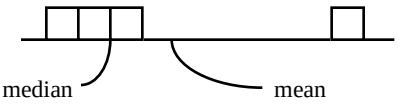
Gould continues:

> "The distribution of variation had to be right skewed, I reasoned. After all, the left of the distribution contains an irrevocable lower boundary of zero (since mesothelioma can only be identified at death or before). Thus there isnt much room for the distribution's lower (or left) half—it must be scrunched up between zero and eight months. But the upper (or right) half can extend out for years and years, even if nobody ultimately survives."

As this probability distribution is skewed (stretched out) to the right, the median is to the left of its mean; Gould's life expectancy is greater than 8 months. (The mean of 20 months suggested in the graph is my ignorant invention.)

The effect of skewness can be seen especially clearly in the case of discrete distributions like the following. Observe that if the right-hand weight is pushed further right the mean will follow, while the median remains between the second and third blocks.

[14] See *Discover* magazine, **6** (June 1985), pp. 40-42.



4.5 Variance; Chebyshev-Markov Inequalities

The **variance**, defined as

$$var(X) = E[(X - E(X))^2],$$

is one of the two commonly used measures of your uncertainty about the value of X. The other is the **"standard deviation"**: $\sigma(X) = \sqrt{var(X)}$.

The formula for the variance can be simplified as follows:[15]

$$var(X) = E(X^2) - (E(X))^2 \tag{1}$$

It is important to note that, in contrast to the linearity of expectations, $E(aX + b) = aE(X) + b$, for variances we have a nonlinear equation:

$$var(aX + b) = a^2\, var(X). \tag{2}$$

Note, too, that in contrast to the unrestricted additivity of E, for var we have

$$var(\sum_{i=1}^{n} I_{H_i}) = \sum_{i=1}^{n} var(I_{H_i}) \text{ if } pr(H_iH_j) = pr(H_i)pr(H_j) \text{ when } i \neq j. \tag{3}$$[16]

This formula can be extended from indicators to general random variables if the requirement of pairwise probabilistic independence, $pr(H_iH_j) = pr(H_i)pr(H_j)$ when $i \neq j$, is replaced by a weaker, analogous requirement on the expectations:[17]

$$var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} var(X_i) \text{ if } E(X_iX_j) = E(X_i)E(X_j) \text{ when } i \neq j. \tag{4}$$

[15] *Proof.* $(X - E(X))^2 = X^2 - 2E(X)X + (E(X))^2$, where $E(X)$ and its square are constants. By linearity of E your expectation of this sum $= E(X^2) - (E(X))^2$.

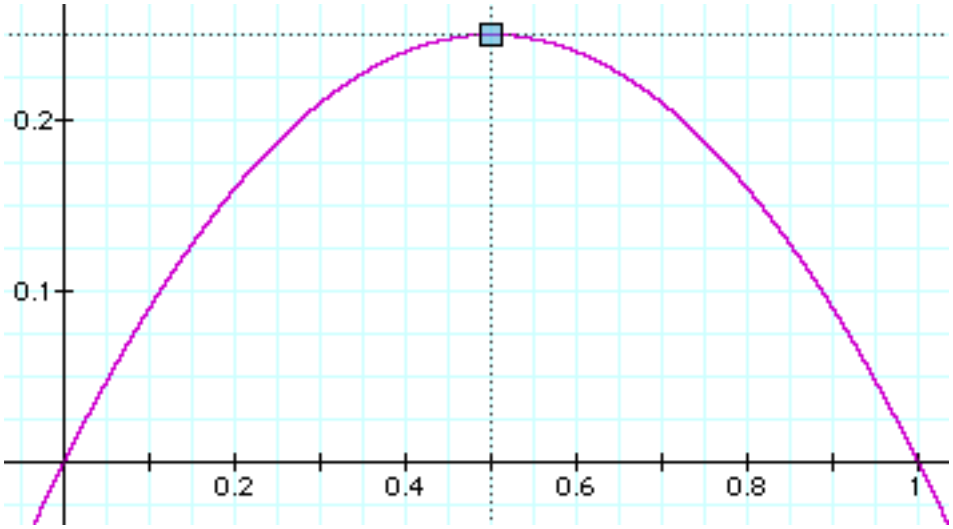
[16] To prove this, use the facts that $(E(I_H))^2 = E(I_H)$, that $E(H_iH_j) = pr(H_i \wedge H_j) = pr(H_i)pr(H_j)$, and that $(\sum_{i=1}^{n} I_{H_i})^2 = \sum_{i=1}^{n}(I_{H_i})^2 + \sum_{i\neq j=1}^{n}(I_{H_i}I_{H_j})$.

[17] *Proof.* By (1), $var(\sum_i X_i) = E((\sum_i X_i)^2) - (E(\sum_i X_i))^2$. This $= \sum_i E((\sum_i X_i)^2) + \sum_{i\neq j=1} E(X_i)E(X_j) - \sum_i (E(X_i))^2 - \sum_{i\neq j=1} E(X_i)E(X_j)$. By (1) this $= \sum_i var(X_i)$.

Now consider the case where X is an indicator, $X = I_H$. Here we have

$$var(I_H) = pr(H)pr(\neg H) \leq \tfrac{1}{4}. \tag{5}$$

Proof. The inequality $pr(H)pr(\neg H) \leq \frac{1}{4}$ follows from the fact that $p(1-p)$ reaches its maximum value of 0.25 when $p = 0.5$, as shown in the graph:



The equality $var(I_H) = pr(H)pr(\neg H)$ follows from (1) because $(I_H)^2 = I_H$ (since I_H can only assume the values 0 and 1) and $E(I_H) = pr(H)$.

There is a simple way of getting an upper bound on your probabilities that non-negative random variables X for which you have finite expectations are at least ϵ away from 0:[18]

Markov's inequality. $pr(X \geq \epsilon) \leq \dfrac{E(X)}{\epsilon}$ if $\epsilon > 0$ and $X \geq 0$.

Proof. If $Y = \{^{\epsilon \text{ when } X \geq \epsilon}_{0 \text{ when } X < \epsilon}$ then by the law of total expectation, $E(Y) =$

$\epsilon \cdot pr(X \geq \epsilon) + 0 \cdot pr(X < \epsilon) = \epsilon \cdot pr(X \geq \epsilon)$, so that $pr(X \geq \epsilon) = \dfrac{E(Y)}{\epsilon}$.

Now since $Y \leq X$ both when $X \geq \epsilon$ and when $X < \epsilon$ we have $E(Y) \leq E(X)$

[18]Here ϵ is a positive real number, not an infinitesmal. The choice of 'ϵ' rather than (say) 'r' just indicates our special interest in the behavior of the inequality as $\epsilon \to 0$.

and, so, Markov's inequality.

Since squares are never negative, the Markov inequality implies that

(6) $$pr(X^2 \geq \epsilon) \leq E(X^2)/\epsilon \text{ if } \epsilon > 0;$$

and from this, since $|X| \geq \epsilon$ if and only if $X^2 \geq \epsilon^2$, we have

(7) $$pr(|X| \geq \epsilon) \leq E(X^2)/\epsilon^2 \text{ if } \epsilon > 0.$$

Now from (7), with '$X - E(X)$' for 'X', we have

$$\textbf{Chebyshev's inequality: } pr(|X - E(X)| \geq \epsilon) \leq \frac{var(X)}{\epsilon^2} \text{ if } \epsilon > 0.$$

Then if your $var(X)$ is small, so is your probability that your expectation of X is far from the truth. When your variance actually vanishes we have the

$$\textbf{Corollary: } var(X) = 0 \text{ if and only if your}$$
$$pr(|X - E(X)| \geq \epsilon) = 0 \text{ for all positive } \epsilon.$$

It is when $var(X) = 0$ that you see X as a constant. In such cases you are quite sure that $X = E(X)$: your probability for that is 1.

4.6 Two Easy Laws of Large Numbers

Laws of large numbers say you how you will expect the "sample" average of the actual values of a number of random variables to behave as that number increases without bound:

$$\textbf{"Sample average"}, \quad S_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

Here we consider two cases—two sorts of random variables—for which such laws have short, easy proofs.

4.6.1 Uncorrelated Random Variables

In this case you see the random variables $X_1, \ldots, X_n$ as

$$\textbf{Uncorrelated: } E(X_i X_j) = E(X_i)E(X_j) \text{ when } i \neq j,$$

and as having variances all of which are bounded above by the same real number, b. Under these conditions there will be a finite upper bound $b/\epsilon^2 n$ on your probability that the sample average will differ by more than ϵ from your expectation of that average. The bound b is independent of the sample size, n, so that for large enough samples your probability that the error $|S_n - E(S_n)|$ in your expectation is ϵ or more gets as small as you please.

(8) Law of Large Numbers for uncorrelated random variables X_i whose variances have a common finite upper bound, $var(X_i) \leq b$:

For any $\epsilon > 0$, $\quad pr(|S_n - E(S_n)| \geq \epsilon) \leq \dfrac{b}{n\epsilon^2}$.

Proof: By (4), $var(S_n) = \dfrac{1}{n^2}\sum_{i=1}^{n} var(X_i)$. Since each $var(X_i)$ is $\leq b$, this is $\leq \dfrac{1}{n^2}(nb) = \dfrac{b}{n}$. Now setting $X = S_n$ in Chebyshev's inequality, we have (8).

"Bernoulli trials" are sequences of propositions H_i (say, "head on the i'th toss") where for some p (=1/2 for coin-tossing) your probability for truth of any n distinct H_i (with $n = 1, 2, \ldots$) is p^n. Corollary (9), below, applies to Bernoulli trials, and also more generally.

Note that if the X_i are indicators of propositions H_i, sample averages S_n are relative frequencies of truth:

$$S_n = \frac{1}{n}\sum_{i=1}^{n} I_{H_i} = \frac{1}{n}(\text{the number of truths among } H_1, \ldots, H_n).$$

Since $E(X_i) = E(I_{H_i}) = pr(H_i) = p$, linearity of E says that $E(S_n) = \frac{1}{n} \cdot np = p$. Now by (4) we can set $b = \frac{1}{4}$ in (8) to get (9):

(9) **Corollary.** If $pr(H_i) = p$ and $pr(H_i \wedge H_j) = p^2$ for all i and $j \neq i$:

For any $\epsilon > 0$, $\quad pr(|S_n - p| \geq \epsilon) \leq \dfrac{1}{4n\epsilon^2}$.

EXAMPLE, **Tossing a fair coin.** If you think there will be at least $n = 25,000$ trials, your probability that the success rate will be within $\epsilon = .01$ of p will fall short of 1 by at most $1/4n\epsilon^2 = .10$.

4.6.2 A Broader Law, Covering some Correlated R.V.'s

For random variables that you see as correlated, your $pr(|S_n - E(S_n)| \geq \epsilon)$ need not tend to 0 as $n \to \infty$. Here the appropriate laws of large numbers will assert that under certain conditions your probabilities $pr(|S_n - S_{n+m}| \geq \epsilon)$ that different sample averages differ by as much as ϵ will tend to 0 as n and m increase without bound.[19]

(10)

Law of Large numbers for random variables $X_1, \ldots, X_n, \ldots X_{n+m}$ where, for all i, j, $E(X_i X_j) = a_1$ if $i = j$ and $E(X_i X_j) = a_2$ if $i \neq j$:

If $0 < \epsilon \leq 1$, then $pr(|S_n - S_{n+m}|) \geq \epsilon) \leq (\frac{1}{n} - \frac{1}{n+m})\frac{a_1 - a_2}{\epsilon^2}$.

If the X_i are pairwise uncorrelated, this holds with $a_1 - a_2 = var(X_i)$.

Proof. By Markov's inequality $pr(|S_n - S_{n+m}| \geq \epsilon) \leq \frac{E(|S_n - S_{n+m}|)}{\epsilon}$, or, equivalently, $pr((S_n - S_{n+m})^2 \geq \epsilon^2) \leq \frac{E(S_n - S_{n+m})^2}{\epsilon^2}$. Now (10) follows from the following

Lemma: $E(S_n - S_{n+m})^2 = (\frac{1}{n} - \frac{1}{n+m})(a_1 - a_2)$

Proof of the lemma requires crunching through some high-school algebra:

(a) $E(S_n - S_{n+m})^2 = E(S_n^2) - 2E(S_n S_{n+m}) + E(S_{n+m}^2)$

(b) $= \frac{E(X_1 + \ldots + X_n)^2}{n^2} - 2\frac{E[(X_1 + \ldots + X_n)(X_1 + \ldots + X_{n+m})]}{n(n+m)} +$ $\frac{E(X_1 + \ldots + X_{n+m})^2}{(n+m)^2}$. Now evaluate the three terms in (b) as in (c)-(e) below. The rationale is essentially the same in all three cases. In case (c) it amounts to the observation that, multiplied out, $(X_1 + \ldots + X_n)^2$ is a sum of n^2 terms $X_i X_j$, where $i = j$ in n of them and $i \neq j$ in the other $n^2 - n$:

(c) $\frac{E(X_1 + \ldots + X_n)^2}{n^2} = \frac{na_1 + n(n-1)a_2}{n^2} = \frac{a_1 - a_2}{n} + a_2.$

In case (d), $(X_1 + \ldots + X_{n+m})^2$ is a sum of $n(n+m)$ terms $X_i X_j$, where $i = j$

[19] See de Finetti, chapter 3 of 'Foresight...' (cited in sec. 1.7, above) or *Theory of Probability,* volume 2 (Wiley, 1975), pp. 215-6.

in n of them and $i \neq j$ in the other $n(n+m) - n$:

(d) $$-\frac{2E[(X_1 + \ldots + X_n)(X_1 + \ldots + X_{n+m})]}{n(n+m)} = -2\frac{na_1 + (n(n+m) - n)a_2}{n(n+m)}$$
$$= -2(\frac{a_1 - a_2}{n+m} + a_2).$$

And in case (e), $(X_1 + \ldots + X_{n+m})^2$ is a sum of $(n+m)^2$ terms X_iX_j, where $i = j$ in $n+m$ of them and $i \neq j$ in the other $(n+m)^2 - (n+m)$:

(e) $$\frac{E(X_1 + \ldots + X_{n+m})^2}{(n+m)^2} = \frac{(n+m)a_1 + (n+m)(n+m-1)a_2}{(n+m)^2} =$$
$$\frac{a_1 - a_2}{n+m} + a_2.$$

(f) Now $E(S_n - S_{n+m})^2 = (c) + (d) + (e) = (\frac{1}{n} - \frac{1}{n+m})(a_1 - a_2)$,

which was to be demonstrated.

If $X_i = I_{H_i}$ in (10), as in coin-tossing, $a_1 = E(X_i^2) = pr(H_i^2) = pr(H_i)$ and (if $i \neq j$) $a_2 = E(X_iX_j) = pr(H_i \wedge H_j)$. Then in the case of constant probability and pairwise independence, $a_1 = p$ and $a_2 = p^2$ and (10) becomes

(11)

Corollary. If $pr(H_i) = p$ and $p(H_i \wedge H_j) = p^2$ for all i and $j \neq i$:
For all $\epsilon > 0$, $pr(|S_n - S_{n+m}|) \geq \epsilon) \leq (\frac{1}{n} - \frac{1}{n+m})\frac{p(1-p)}{\epsilon^2}$.

4.6.3 Covariance and Noncorrelation

The **covariance** of two random variables is defined:

(12) $$cov(X, Y) = E([X - E(X)][Y - E(Y)]).$$

As is easily verified, this is equivalent to

(13) $$cov(X, Y) = E(XY) - E(X)E(Y).$$

Then non-correlation is zero covariance; and $cov(X, X) = var(X)$ while $cov(X, -X) = -var(X)$.

A useful variant of covariance is the **correlation coefficient**:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)} \tag{14}$$

Then the random variables X, Y which you see as uncorrelated are those for which your $\rho(X,Y) = 0$; and $\rho(X,X) = 1$ while $\rho(X,-X) = -1$. It turns out that

$$\boxed{-1 \leq \rho(X,Y) \leq 1, \text{ where the extreme values are assumed if and only for particular constants } a, b \text{ you are certain that } Y = aX + b.} \tag{15}$$

Proof: see problems xx-yy in sec. zz.

Even the minimally nonlinear dependency of X^2 on X noted in example 2 was too complex to make the correlation coefficient go positive. 'Correlation' in the technical sense is short for 'linear correlation'.