CHAPTER 9

REPETITIVE EXPERIMENTS – PROBABILITY AND FREQUENCY

> "*The essence of the present theory is that no probability, direct, prior, or posterior, is simply a frequency.*" — H. Jeffreys (1939)

We have developed probability theory as a generalized logic of plausible inference which should apply, in principle, to any situation where we do not have enough information to permit deductive reasoning. We have seen it applied successfully in simple prototype examples of nearly all the current problems of inference, including sampling theory, hypothesis testing, and parameter estimation.

However, most of probability theory as treated in the past 100 years has confined attention to a special case of this, in which one tries to predict the results of, or draw inferences from, some experiment that can be repeated indefinitely under what appear to be identical conditions; but which nevertheless persists in giving different results on different trials. Indeed, virtually all application–oriented expositions *define* probability as meaning 'limiting frequency in independent repetitions of a random experiment' rather than as an element of logic. The mathematically oriented often define it more abstractly, merely as an additive measure without any specific connection to the real world. However, when they turn to applications, they too tend to think of probability in terms of frequency. It is important that we understand the exact relation between these conventional treatments and the theory being developed here.

Some of these relations have been seen already; in the last five Chapters we have shown that probability theory as logic can be applied consistently in many problems of inference that do not fit into the frequentist preconceptions, and so would be considered beyond the scope of probability theory. Evidently, the problems that can be solved by frequentist probability theory form a subclass of those that are amenable to logical probability theory, but it is not yet clear just what that subclass is. In the present Chapter we seek to clarify this with some surprising results, including a new understanding of the role of induction in science.

There are also many problems where the attempt to use frequentist probability theory in inference leads to nonsense or disaster. We postpone examination of this pathology to later Chapters, particularly Chapter 17.

Physical Experiments

Our first example of such a repetitive experiment appeared in Chapter 3, where we considered sampling with replacement from an urn, and noted that even there great complications arise. But we managed to muddle our way through them by the conceptual device of "randomization" which, although ill–defined, had enough intuitive force to overcome the fundamental lack of logical justification.

Now we want to consider general repetitive experiments where there need not be any resemblance to drawing from an urn, and for which those complications may be far greater and more diverse than they were for the urn. But at least we know that any such experiment is subject to physical law. If it consists of tossing a coin or die, it will surely conform to the laws of Newtonian mechanics, well known for 300 years. If it consists of giving a new medicine to a variety of patients, the principles of biochemistry and physiology, only partially understood at present, surely determine the possible effects that can be observed. An experiment in high–energy elementary particle physics is subject to physical laws about which we are about equally ignorant; but even here well–established general principles (conservation of charge, angular momentum, *etc.*) restrict the possibilities.

Clearly, competent inferences about any such experiment must take into account whatever is presently known concerning the physical laws that apply to the situation. Generally, this knowledge will determine the "model" that we prescribe in the statement of the problem. If one fails to take account of real physical situation and the known physical laws that apply, then the most impeccably rigorous mathematics from that point on will not guard against producing nonsense or worse. The literature gives much testimony to this.

In any repeatable experiment or measurement, some relevant factors are the same at each trial (whether or not the experimenter is consciously trying to hold them constant – or is even consciously aware of them), and some vary in a way not under the control of the experimenter. Those that are the same (whether from the experimenter's good control of conditions or from his failure to influence them at all) are called *systematic*. Those which vary in an uncontrolled way are often called *random*, a term which we shall avoid for the present, because in current English usage it carries some very wrong connotations.[†]

In this Chapter we examine in detail how our robot reasons about a repetitive experiment. Our aim is to find the logical relations between the information it has and the kind of predictions it is able to make. Let our experiment consist of n trials, with m possible results at each trial; if it consists of tossing a coin, then $m = 2$; for a die, $m = 6$. If we are administering a vaccine to a sequence of patients, then m is the number of distinguishable reactions to the treatment, n is the number of patients, *etc.*

At this point one would say, conventionally, something like: "Each trial is capable of giving any one of m possible results, so in n trials there are $N = m^n$ different conceivable outcomes." However, the exact meaning of this is not clear: is it a statement or assumption of physical fact, or only a description of the robot's information? The content and range of validity of what we are doing depends on the answer.

The number m may be regarded, always, as a description of the state of knowledge in which we conduct a probability analysis; but this may or may not correspond to the number of real possibilities actually existing in Nature. On examining a cubical die, we feel rather confident in taking $m = 6$; but in general we cannot know in advance, with certainty, how many different results are possible. Some of the most important problems of inference are of the "Charles Darwin" type:

Exercise 9.1: When Charles Darwin first landed on the Galapagos Islands in September 1835, he had no idea how many different species of plants he would find there. Having examined $n = 122$ specimens and finding that they can be classified into $m = 19$ different species, what is the probability that there are still more species, as yet unobserved? At what point does one decide to stop collecting specimens because it is unlikely that anything more will be learned? This problem is much like that of the sequential test of Chapter 4, although we are now asking a different question. It requires judgment about the real world in setting up the mathematical model (that is, in the prior information used in choosing the appropriate hypothesis space), but the final conclusions are quite insensitive to the exact choice made, so persons with reasonable judgment will be led to substantially the same conclusions.

[†] To many, the term "random" signifies on the one hand lack of physical determination of the individual results, *but at the same time*, operation of a physically real 'propensity' rigidly fixing long–run frequencies. Naturally, such a self–contradictory view of things gives rise to endless conceptual difficulties and confusion, throughout the literature of every field that uses probability theory. We note some typical examples in Chapter 10, where we confront this idea of 'randomness' with the laws of physics.

In general, then, far from being a known physical fact, the number m should be understood to be simply the number of known results per trial *that we shall take into account in the present calculation.* But the very purpose of the calculation may be to learn how m is related to the true number of possibilities existing in Nature. Then it is perhaps being stated most defensibly if we say that when we specify m we are defining *a tentative working hypothesis*, whose consequences we want to learn.

For clarity, we use the word "result" for a single trial, while "outcome" refers to the experiment as a whole. Thus one outcome consists of the enumeration of n results (including their order if the experiment is conducted in such a way that an ordering is defined and known). Then we may say that the number of outcomes *being considered in the present calculation* is $N = m^n$.

Denote the result of the k'th trial by $r_k, (1 \le r_k \le m,\ 1 \le k \le n)$. Then any outcome of the experiment can be indicated by specifying the numbers $\{r_1, \ldots, r_n\}$, which constitute a conceivable data set D. Since the different outcomes are mutually exclusive and exhaustive, if our robot is given any information I about the experiment, the most general probability assignment it can make is a set of non–negative real numbers

$$P(D|I) = f(r_1 \ldots r_n) \tag{9–1}$$

satisfying

$$\sum_{r_1=1}^{m} \sum_{r_2=1}^{m} \cdots \sum_{r_n=1}^{m} f(r_1 \ldots r_n) = 1\ . \tag{9–2}$$

Note, as a convenience, that we may regard the numbers r_k as digits (*modulo* m) in a number R expressed in the base m number system; $0 \le R \le N - 1$. Since our robot, however poorly informed it may be about the real world, is an accomplished manipulator of numbers, we may instruct it to communicate with us in the base m number system instead in the decimal (base 10) number system that you and I have been trained to use because of an anatomical peculiarity of humans.

For example, suppose that our experiment consists of tossing a die four times; there are $m = 6$ possible results at each trial, and $N = 6^4 = 1296$ possible outcomes for the experiment. Then to indicate the outcome that is designated number 836 in the decimal system, the robot notes that

$$836 = (3 \times 6^3) + (5 \times 6^2) + (1 \times 6^1) + (2 \times 6^0)$$

and so, in the base 6 system the robot displays this as outcome number 3512.

But unknown to the robot, this has a deeper meaning to you and me; for us, this represents the outcome in which the first toss gave three spots up, the second gave five spots, the third gave one spot, and the fourth toss gave two spots (since in the base 6 system the individual digits r_k have meaning only *modulo* 6, the display $5024 = 5624$ represents an outcome in which the second toss yielded six spots up).

More generally, for an experiment with m possible results at each trial, repeated n times, we communicate in the base m number system, whereupon each number displayed will have exactly n digits, and for us the k'th digit will represent, *mod* m, the result of the k'th trial. By this device we trick our robot into taking instructions and giving its conclusions in a format which has for us an entirely different meaning. We can now ask the robot for its predictions on any question we care to ask about the digits in the display number, and this will never betray to the robot that it is really making predictions about a repetitive physical experiment (for the robot, by construction as discussed in Chapter 4, always accepts what we tell it as the literal truth).

With the conceptual problem defined as carefully as we know how to do, we may turn finally to the actual calculations. We noted in the discussion following Eq. (2–65) that, depending on details of the information I, many different probability assignments (9–1) might be appropriate; consider first the obvious simplest case of all.

The Poorly Informed Robot

Suppose we tell the robot only that there are N possibilities, and give no other information. That is, the robot is not only ignorant about the relevant physical laws; it is not even told that the full experiment consists of n repetitions of a simpler one. For it, the situation is as if there were only a single trial, with N possible results, the "mechanism" being completely unknown.

At this point, you might object that we have withheld from the robot some very important information, that must be of crucial importance for any rational inferences about the experiment; and so we have. Nevertheless, it is important that we understand the surprising consequences of neglecting that information.

But what meaningful predictions about the experiment could the robot possibly make, when it is in such a primitive state of ignorance that it does not even know that there is any repetitive experiment involved? Actually, the poorly informed robot is far from helpless; although it is hopelessly naïve in some respects, nevertheless it is already able to make a surprisingly large number of correct predictions for purely combinatorial reasons (this should give us some respect for the cogency of multiplicity factors, which can mask a lot of ignorance).

Let us see first just what those poorly informed predictions are; then we can give the robot additional pertinent pieces of information and see how its predictions are revised as it comes to know more and more about the real physical experiment. In this way we can follow the robot's education step by step, until it reaches a level of sophistication comparable to (in many cases, exceeding) that displayed by real scientists and statisticians discussing real experiments.

Denote this initial state of ignorance (the robot knows only the number N of possible outcomes and nothing else) by I_0. The principle of indifference (2–74) then applies; the robot's "sample space" or "hypothesis space" consists of $N = m^n$ discrete points, and to each it assigns probability N^{-1}. Any proposition A that is defined to be true on a subset containing $M(A)$ points and false on the rest will, by the rule (2–76), then be assigned the probability

$$P(A|I_0) = \frac{M(A)}{N} \ , \tag{9–3}$$

just the *frequency* with which A is true on the full set. This trivial–looking result summarizes everything the robot can say on the prior information I_0, and it illustrates again that connections between probability and frequency appear automatically in probability theory as logic, as mathematical consequences of the rules, whenever they are relevant to the problem.

Consider n tosses of a die, $m = 6$; the probability (9–1) of any completely specified outcome is

$$f(r_1 \ldots r_n|I_0) = \frac{1}{6^n}, \qquad 1 \le r_k \le 6, \quad 1 \le k \le \ n \ . \tag{9–4}$$

What is the probability that the first toss gives three spots, regardless of what happens later? We ask the robot for the probability that the first digit $r_1 = 3$. Then the 6^{n-1} propositions

$$A(r_2 \ldots r_n) \equiv \text{“}r_1 = 3 \text{ and the remaining digits are } r_2 \ldots r_n\text{”}$$

are mutually exclusive, and so (2–64) applies:

$$P(r_1 = 3|I_0) = \sum_{r_2=1}^{6} \cdots \sum_{r_n=1}^{6} f(3, r_2 \ldots r_n|I_0) = 6^{n-1}\, f(r_1 \ldots r_n|I_0) = \frac{1}{6} \tag{9–5}$$

[Note that the statement "$r_1 = 3$" is a proposition, so by our notational rules in Appendix B we are allowed to put it in a formal probability symbol.]

But by symmetry, if we had asked for the probability that any specified (k'th) toss gives any specified (i'th) result, the calculation would have been the same:

$$P(r_k = i|I_0) = \frac{1}{6}\,, \qquad 1 \le i \le 6, \quad 1 \le k \le n\,. \tag{9–6}$$

Now, what is the probability that the first toss gives i spots, and the second gives j spots? The robot's calculation is just like the above; the results of the remaining tosses comprise 6^{n-2} mutually exclusive possibilities, and so

$$\begin{aligned} P(r_1 = i,\ r_2 = j|I_0) &= \sum_{r_3=1}^{6} \cdots \sum_{r_n=1}^{6} f(i, j, r_3 \ldots r_n|I_0) = \ 6^{n-2}\ f(r_1 \ldots r_n|I_0) = \frac{1}{6^2} \\ &= \frac{1}{36} \end{aligned} \tag{9–7}$$

and by symmetry the answer would have been the same for any two different tosses. Similarly, the robot will tell us that the probability of any specified outcomes at any three different tosses is

$$f(r_i\ r_j\ r_k|I_0) \ = \ \frac{1}{6^3} \ = \ \frac{1}{216} \tag{9–8}$$

and so on!

Let us now try to educate the robot. Suppose we give it the additional information that, to you and me, means that the first toss gave 3 spots. But we tell this to the robot in the form: out of the originally possible N outcomes, the correct one belongs to the subclass for which the first digit is $r_1 = 3$. With this additional information, what probability will it now assign to the proposition $r_2 = j$? This conditional probability is determined by the product rule (2–46):

$$f(r_2|r_1 I_0) \ = \ \frac{f(r_1 r_2|I_0)}{f(r_1|I_0)} \tag{9–9}$$

or, using (9–6), (9–7),

$$f(r_2|r_1 I_0) = \frac{1/36}{1/6} = \frac{1}{6} = f(r_2|I_0)\,. \tag{9–10}$$

The robot's prediction is unchanged. If we tell it the result of the first two tosses and ask for its predictions about the third, we have from (9–8) the same result:

$$f(r_3|r_1 r_2 I_0) = \frac{f(r_3 r_1 r_2|I_0)}{f(r_1 r_2|I_0)} = \frac{1/216}{1/36} = \frac{1}{6} = f(r_3|I_0)\,. \tag{9–11}$$

We can continue in this way, and will find that if we tell the robot the results of any number of tosses, this will have no effect at all on its predictions for the remaining ones.

It appears that the robot is in such a profound state of ignorance I_0 that it cannot be educated. However, if it does not respond to one kind of instruction, perhaps it will respond to another. But first we need to understand the cause of the difficulty.

Induction

In what way does the robot's behavior surprise us? Its reasoning here is different from the way you and I would reason, in that the robot does not seem to learn from the past. If we were told that the first dozen digits were all 3, you and I would take the hint and start placing our bets on 3 for the next digit. But the poorly informed robot does not take the hint, no matter how many times it is given.

More generally, if you or I could perceive any regular pattern in the previous results, we would more or less expect it to continue; this is the reasoning process called "induction". The robot does not yet see how to reason inductively. However, the robot must do all things quantitatively, and you and I would have to admit that we are not certain whether the regularity will continue. It only seems somewhat likely, but our intuition does not tell us how likely. So our intuition, again, gives us only a qualitative "sense of direction" in which we feel the robot's quantitative reasoning ought to go.

Note that what we are calling "induction" is a very different process from what is called, confusingly, "mathematical induction". The latter is a rigorous deductive process, and we are not concerned with it here.

The problem of "justifying induction" has been a difficult one for the conventional formulations of probability theory usually taught to scientists, and the nemesis of some philosophers. For example, the philosopher Karl Popper (1974) has gone so far as to flatly deny the possibility of induction. He asked the rhetorical question: "*Are we rationally justified in reasoning from repeated instances of which we have experience to instances of which we have no experience*?" This is, quite literally, the poorly informed robot speaking to us, and wanting us to answer "*No*!". But we want to show that a better informed robot will answer: "*Yes, if we have prior information connecting the different trials*" and give specific circumstances that enable induction to be made.

The difficulty has seemed particularly acute in the theory of survey sampling, which corresponds closely to our equations above. Having questioned 1000 people and found that 672 of them favor proposition A in the next election, by what right do the pollsters jump to the conclusion that about 67 ± 3 percent of the millions not surveyed also favor proposition A? For the poorly informed robot (and, apparently, for Popper too), learning the opinions of any number of persons tells it nothing about the opinions of anyone else.

The same logical problem appears in many other situations. In physics, suppose we measured the energies of 1000 atoms, and found that 672 of them were in excited states, the rest in the ground state. Do we have any right to conclude that about 67 percent of the 10^{23} other atoms not measured are also in excited states? Or, 1000 cancer patients were given a new treatment and 672 of them recovered; then in what sense is one justified in predicting that this treatment will also lead to recovery in about 67% of future patients? On prior information I_0 there is no justification at all for such inferences.

As these examples show, the problem of logical justification of induction (*i.e.*, of clarifying the exact meaning of the statements, and the exact sense in which they can be supported by logical analysis) is important as well as difficult. We hope to show that only probability theory as logic can solve this problem.

Are There General Inductive Rules?

What is shown by (9–10) and (9–11) is that on the information I_0 the results of different tosses are, logically, completely independent propositions; giving the robot any information whatsoever about the results of specified tosses, tells it nothing relevant to any other toss. The reason for this was stressed above: the robot does not yet know that the successive digits $\{r_1, r_2 \ldots\}$ represent successive repetitions of the *same* experiment. It can be educated out of this state only by giving it some kind of information that has relevance to all tosses; for example, if we tell it something, however slight, about some property that is common to all trials.

Perhaps, then, we might learn by introspection: what is that extra "hidden" information, common to all trials, that you and I are using, unconsciously, when we do inductive reasoning? Then we might try giving this hidden information to the robot (*i.e.*, incorporate it into our equations).

But a very little introspection is enough to make us aware that there is no one piece of hidden information; there are many different kinds. Indeed, the inductive reasoning that we all do varies widely, even for identical data, as our prior knowledge about the experiment varies. Sometimes we "take the hint" immediately, and sometimes we are as slow to do it as the poorly informed robot.

For example, suppose the data are that the first three tosses of a coin have all yielded "heads": $D = H_1 H_2 H_3$. What is our intuitive probability $P(H_4|DI)$ for heads on the fourth toss? This depends very much on what that prior information I is. On prior information I_0 the answer is always $p(H_4|DI_0) = 1/2$, whatever the data. Two other possibilities are:

$I_1 \equiv$ "We have been allowed to examine the coin carefully and observe the tossing. We know that the coin has a head and a tail and is perfectly symmetrical, with its center of gravity in the right place, and we saw nothing peculiar in the way it was tossed."

$I_2 \equiv$ "We were not allowed to examine the coin, and we are very dubious about the 'honesty' of either the coin or the tosser."

On information I_1, our intuition will probably tell us that the prior evidence of the symmetry of the coin far outweighs the evidence of three tosses; so we shall ignore the data and again assign $P(H_4|DI_1) = 1/2$.

But on information I_2 we would consider the data to have some cogency: we would feel that the fact of three heads and no tails constitutes some evidence (although certainly not proof) that some systematic influence is at work favoring heads, and so we would assign $P(H_4|DI_2) > 1/2$. Then we would be doing real inductive reasoning.

But now we seem to be facing a paradox. For I_1 represents a great deal more information than does I_2; yet it is $P(H_4|DI_1)$ that agrees with the poorly informed robot! In fact, it is easy to see that all our inferences based on I_1 agree with those of the poorly informed robot, as long as the prior evidence of symmetry outweighs the evidence of the data).

However, this is only an example of something that we have surely noted many times in other contexts. The fact that one person has far greater knowledge than another does not mean that they necessarily disagree; an idiot might guess the same truth that a scholar has spent years establishing. All the same, it does call for some deep thought to understand why knowledge of perfect symmetry could leave us making the same inferences as does the poorly informed robot.

As a start on this, note that we would not be able to assign any definite numerical value to $P(H_4|DI_2)$ until that vague information I_2 is specified much more clearly. For example, consider the extreme case:

$I_3 \equiv$ "We know that the coin is a trick one, that has either two heads or two tails; but we do not know which."

Then we would, of course, assign $P(H_4|DI_3) = 1$; in this state of prior knowledge, the evidence of a single toss is already conclusive. It is not possible to take the hint any more strongly than this.

As a second clue, note that our robot did seem, at first glance, to be doing inductive reasoning of a kind back in Chapter 3, for example in (3–13), where we examined the hypergeometric distribution. But on second glance it was doing "reverse induction"; the more red balls had been drawn, the lower its probability for red in the future. And this reverse induction disappeared when we went on to the limit of the binomial distribution.

But you and I could also be persuaded to do reverse induction in coin tossing. Consider the prior information:

$I_4 \equiv$ "The coin has a concealed inner mechanism that constrains it to give exactly 50 heads and 50 tails in the next 100 tosses"

On this prior information, we would say that tossing the coin is, for the next 100 times, equivalent to drawing from an urn that contains initially 50 red balls and 50 white ones. We could then use the product rule as in (9–9) but with the hypergeometric distribution $h(r|N, M, n)$ of (3–18):

$$P(H_4|DI_4) = \frac{h(4|100, 50, 4)}{h(3|100, 50, 3)} = \frac{0.05873}{0.12121} = 0.4845 < \frac{1}{2}$$

But in this case it is easier to reason it out directly: $P(H_4|DI_4) = (M - 3)/(N - 3) = 47/97 = 0.4845$.

The great variety of different results that we have found from the same data makes it clear that there can be no such thing as a single universal inductive rule and, in view of the unlimited variety of different kinds of conceivable prior information, makes it seem dubious that there could exist even a classification of all inductive rules by any system of parameters.

Nevertheless, such a classification was attempted by the philosopher R. Carnap (1891–1970), who found (Carnap, 1952) a continuum of rules determined by a single parameter λ, $(0 < \lambda < \infty)$. But ironically, Carnap's rules turned out to be identical with those given, on the basis of entirely different reasoning, by Laplace in the 18'th Century (the "rule of succession" and its generalizations) that had been rejected as metaphysical nonsense by statisticians and philosophers.[†]

Laplace was not considering the general problem of induction, but was only finding the consequences of a certain type of prior information, so the fact that he did not obtain every conceivable inductive rule never arose and would have been of no concern to him. In the meantime, superior analyses of Laplace's problem had been given by W. E. Johnson (1932), Bruno de Finetti (1937) and Harold Jeffreys (1939), of which Carnap seemed unaware.

Carnap is seeking the general inductive rule (*i.e.*, the rule by which, given the record of past results, one can make the best possible prediction of future ones). But his exposition wanders off into abstract symbolic logic without ever considering a specific real example; and so it never rises to the level of seeing that *different inductive rules correspond to different prior information.* It seems to us obvious, from arguments like the above, that this is the primary fact controlling induction, without which the problem cannot even be stated, much less solved. Yet neither the term "prior information" nor the concept ever appears in Carnap's exposition.

This should give a good idea of the level of confusion that exists in this field, and the reason for it; conventional frequentist probability theory simply ignores prior information[‡] and – just for that reason – it is helpless to account for induction. Fortunately, probability theory as logic is able to deal with the full problem. But to show this we need to develop our mathematical techniques somewhat further, in the way that Laplace showed us some 200 years ago.

[†] Carnap (*loc cit*, p. 35), like Venn, claims that Laplace's rule is inconsistent (in spite of the fact that it is identical with his own rule); we examine these claims in Chapter 18 and find, in agreement with R. A. Fisher (1956), that they have misapplied Laplace's rule by ignoring the necessary conditions required for its derivation.

[‡] This is an understatement. Some frequentists take a militant stand *against* prior information, thereby guaranteeing failure in trying to understand induction. We have already seen, in the example of Bertrand at the end of Chapter 6, how disastrously wrong this is in other problems of inference.

Multiplicity Factors

In spite of the formal simplicity of (9–3), the actual numerical evaluation of $P(A|I_0)$ for a complicated proposition A may involve immense combinatorial calculations. For example, suppose we toss a die twelve times. The number of conceivable outcomes is

$$6^{12} = 2.18 \times 10^9 \ ,$$

which is about equal to the number of minutes since the Great Pyramid was built. The geologists and astrophysicists tell us that the age of the universe is about 10^{10} years, or 3×10^{17} seconds. Thus, in thirty tosses of a die, the number of possibilities ($6^{30} = 2.21 \times 10^{23}$) is about equal to the number of microseconds in the age of the universe. Yet we shall be particularly interested in evaluating quantities like (9–3) pertaining to a famous experiment involving 20,000 tosses of a die!

It is true that we are concerned with finite sets; but they can be rather large and we need to learn how to calculate on them. An exact calculation will generally involve intricate number–theoretic details (such as whether n is a prime number, whether it is odd or even, *etc.*), and may require many different analytical expressions for different n; yet in view of the large numbers there will be enormously good approximations which turn out to be easy to calculate.

A large class of problems may be fit into the following scheme. Let $\{g_1, g_2 \dots g_m\}$ be any set of m finite real numbers. For concreteness, one may think of g_i as the "value" or the "gain" of observing the i'th result in any trial (perhaps the number of pennies we win whenever that result occurs), but the following considerations are independent of whatever meaning we attach to the $\{g_j\}$, with the proviso that they are additive; *i.e.*, sums like $g_1 + g_2$ are to be, like sums of pennies, meaningful to us. Or, perhaps g_j is the excitation energy of the j'th atom, in which case G is the total excitation energy of the sampled atoms. Or, perhaps g_j is the size of the j'th account in a bank, in which case G is the total deposits in the accounts inspected. The total amount of G found in the experiment is then

$$G = \sum_{k=1}^{n} g(r_k) = \sum_{j=1}^{m} n_j \, g_j \qquad (9\text{–}12)$$

where the sample number n_j is the number of times the result r_j occurred. If we ask the robot for the probability of obtaining this amount, it will answer, from (9–3),

$$f(G|n, I_0) = \frac{M(n, G)}{N} \qquad (9\text{–}13)$$

where $M(n, G)$ is the multiplicity of the event G; i.e., the number of different outcomes which yield the value G (we now indicate in it also the number of trials n – to the robot, the number of digits which define an outcome – because we want to allow this to vary). Many probabilities are determined by this multiplicity factor; for example, suppose we are told the result of the i'th trial: $r_i = j$, where $1 \le i \le n$, $1 \le j \le m$. Then the total G becomes, in place of (9–12),

$$G = g_j + \sum_{k \neq j} n_k \, g_k \qquad (9\text{–}14)$$

and the multiplicity of this is, evidently, $M(n-1, G-g_j)$. Therefore the probability of getting the total gain G is changed to

$$p(G|r_i = j, n, I_0) = \frac{M(n-1, G-g_j)}{m^{n-1}} \qquad (9\text{–}15)$$

and, given only I_0, the probability of the event $r_i = j$ is, from (9–6),

$$p(r_i = j|n, I_0) = \frac{1}{m} \tag{9–16}$$

This gives us everything we need to apply Bayes' theorem conditional on G:

$$p(r_i = j|G, n, I_0) = p(r_i = j|n, I_0)\,\frac{p(G|r_i = j, n, I_0)}{p(G|n, I_0)} \tag{9–17}$$

or,

$$p(r_i = j|G, n, I_0) = \frac{1}{m}\,\frac{[M(n-1, G-g_j)/m^{n-1}]}{[M(n, G)/m^n]} = \frac{M(n-1, G-g_j)}{M(n, G)} \tag{9–18}$$

Exercise 9.2: Extend this result to find the joint probability

$$p(r_i = j, r_s = t|G, n, I_0) = M(n-2, G-g_j-g_t)/M(n, G) \tag{9–19}$$

as a ratio of multiplicities.

Many problems can be solved if we can calculate the multiplicity factor $M(n, G)$; as noted it may require an immense calculation to find it exactly, but there are relatively simple approximations which become enormously good for large n.

Partition Function Algorithms

Formally, the above multiplicity varies with n and G in a simple way. Expanding $M(n, G)$ according to the result of the n'th trial gives the recursion relation

$$M(n, G) = \sum_{j=1}^{m} M(n-1, G-g_j)\ . \tag{9–20}$$

This is a linear difference equation with constant coefficients in both n and G, so it must have elementary solutions of exponential form:

$$\exp(\alpha n + \lambda G)\ . \tag{9–21}$$

On substitution into (9–19) we find that this is a solution of the difference equation if α and λ are related by

$$e^{\alpha} = Z(\lambda) \equiv \sum_{j=1}^{m} e^{-\lambda g_j}\,. \tag{9–22}$$

The function $Z(\lambda)$ is called the *partition function*, and it will have a fundamental importance throughout all of probability theory. An arbitrary superposition of such elementary solutions:

$$H(n, G) = \int Z^n(\lambda)\, e^{\lambda G}\, h(\lambda)\, d\,\lambda \tag{9–23}$$

is a formal solution of (9–19). However, the true $M(n, G)$ also satisfies the initial condition $M(0, G) = \delta(G, 0)$ and is defined only for certain discrete values of $G = \Sigma n_j g_j$, the values that are possible results of n trials.

Since (9–23) has the form of an inverse Laplace transform, let us note the discrete Laplace transform of $M(n,G)$. Suppose we multiply $M(n,G)$ by $\exp(-\lambda G)$ and sum over all possible values of G. This sum contains a contribution from every possible outcome of the experiment, and so it can be expressed equally well as a sum over all possible sample numbers:

$$\sum_G e^{-\lambda G}\, M(n,G) = \sum_{\{n_j\}} W(n_1 \ldots n_m)\, \exp(-\lambda \Sigma n_j g_j)\,, \tag{9–24}$$

where the multinomial coefficient

$$W(n_1 \ldots n_m) \equiv \frac{N!}{n_1! \ldots n_m!} \tag{9–25}$$

is the number of outcomes that have the sample numbers $\{n_1 \ldots n_m\}$, and we sum over the region $\{R : \sum n_j = N,\ n_j \geq 0\}$. But, comparing with the multinomial expansion of $(x_1 + \cdots + x_m)^n$, this is just

$$\sum_G e^{-\lambda G}\, M(n,G) = Z^n(\lambda)\,. \tag{9–26}$$

Therefore the proper choice of the function $h(\lambda)$ and path of integration in (9–23) is the one that makes (9–23) and (9–26) a Laplace transform pair. To find it, note that the integrand in (9–23) contains a sum of a finite number of terms:

$$Z^n(\lambda)e^{\lambda G} = \sum_k M(n, G_k)e^{\lambda(G-G_k)} \tag{9–27}$$

where $\{G_k\}$ are the possible gains. Therefore it suffices to consider a single term. Now an integral over an infinite domain is by definition the limit of a sequence of integrals over finite domains, so consider the integral

$$I(u) \equiv \frac{1}{2i}\int_{-iu}^{iu} e^{\lambda(G-G_k)}\, d\lambda = \frac{\sin u(G-G_k)}{G-G_k}\,. \tag{9–28}$$

As a function of G, this has a single maximum of height u, width about π/u. In fact, $\int \sin ux/x dx = \pi$ independent of u. As $u \to \infty$, we have $I(u) \to \pi\delta(G - G_k)$, so

$$\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} Z^n(\lambda)\, e^{\lambda G}\, d\lambda = \sum_k M(n, G_k)\, \delta(G - G_k) \tag{9–29}$$

and of course (9–26) can be written more explicitly as

$$Z^n(\lambda) = \int e^{-\lambda G}\, q(G)\, dG \tag{9–30}$$

where

$$q(G) \equiv \sum_k M(n, G_k)\, \delta(G - G_k)\,. \tag{9–31}$$

and so the required result is: $Z^n(\lambda)$ and $q(G)$ are a standard Laplace transform pair.[†]

[†] This illustrates again how awkward it would be to try to conduct substantive analytical work without delta functions; they arise naturally and inevitably in the course of many calculations, and they can be evaded only by elaborate and quite unnecessary subterfuges. The reader is expected to be aware of the work of Lighthill establishing this rigorously, as noted in Appendices B and F.

We consider the use of this presently, but note first that in many cases (9–26) is all we need to solve combinatorial problems.

Equation (9–26) says that the number of ways $M(n, G)$ in which a particular value G can be realized is just the coefficient of $\exp(-\lambda G)$ in $Z^n(\lambda)$; in other words, $Z(\lambda)$ raised to the n'th power displays the exact way in which all the possible outcomes in n trials are partitioned among the possible values of G, which indicates why the name 'partition function" is appropriate.

In some simple problems, this observation gives us the solution by mere inspection of $Z^n(\lambda)$. For example, if we make the choice

$$g_j \equiv \ \delta(i, 1) \tag{9–32}$$

then the total G is just the first sample number:

$$G = \sum n_j g_j = n_1 \ . \tag{9–33}$$

The partition function (9–22) is then

$$Z(\lambda) = e^{-\lambda} + m - 1 \tag{9–34}$$

and from Newton's binomial expansion,

$$Z^n(\lambda) = \sum_{s=0}^{n} \binom{n}{s} \ e^{-\lambda s} \, (m-1)^{n-s} \ . \tag{9–35}$$

$M(n, G) = M(n, n_1)$ is then the coefficient of $\exp(-\lambda n_1)$ in this expression:

$$M(n, G) = M(n, n_1) = \binom{n}{n_1} (m-1)^{n-n_1} \ . \tag{9–36}$$

In this simple case, the counting could have been done also as: $M(n, n_1)$ = (number of ways of choosing n_1 trials out of n) × (number of ways of allocating the remaining $m-1$ trial results to the remaining $n - n_1$ trials). However, the partition function method works just as well in far more complicated problems; and even in this example the partition function method, once understood, is easier to use.

In the choice (9–32) we separated off the trial result $j = 1$ for special attention. More generally, suppose we separate the m trial results arbitrarily into a subset S containing s of them, and the complementary subset $\overline{S}$ consisting of the $(m - s)$ remaining ones, where $1 < s < m$. Call any result in the subset S a "success", any in $\overline{S}$ a "failure". Then we replace (9–32) by

$$g_j = \left\{ \begin{matrix} 1, & j \in S \\ 0, & \text{otherwise} \end{matrix} \right\} \tag{9–37}$$

and Equations (9–33)–(9–36) are generalized as follows. G is now the total number of successes, called traditionally r:

$$G = \sum_{j=1}^{m} n_j \, g_j = r \tag{9–38}$$

which, like n_1, can take on all values in $0 \le r \le n$. The partition function now becomes

$$Z(\lambda) = s \, e^{-\lambda} + m - s \tag{9–39}$$

from which

$$Z^n(\lambda) = \sum_{r=0}^{n} \binom{n}{r} s^r e^{-\lambda r} (m-s)^{n-r} \tag{9–40}$$

and

$$M(n,G) = M(n,r) = \binom{n}{r} \; s^r \; (m-s)^{n-r} \; . \tag{9–41}$$

From (9–13), the poorly informed robot's probability for r successes is therefore

$$P(G = r|I_0) = \binom{n}{r} \, p^r \, (1-p)^{n-r} \; , \qquad 0 \le r \le n \tag{9–42}$$

where, on the right–hand side, $p = s/m$.

But this is just the binomial distribution $b(r|n,p)$, whose derivation cost us so much conceptual agonizing in Chapter 3; now seen in a new light. In Chapter 3, we obtained the binomial distribution (3–74) as the limiting form in drawing from an infinitely large urn, and again as a randomized approximate form (3–79) in drawing with replacement from a finite urn; but in neither case was it exact for a finite urn. Now we have found a case where the binomial distribution arises for a different reason and it is exact for a finite sample space.

This quantitative exactness is a consequence of our making the problem more abstract; there is now, in the prior information I_0, no mention of complicated physical properties such as those of urns, balls, and hands reaching in. But more important, and surprising, is simply the qualitative fact that the binomial distribution, ostensibly arising out of repeated sampling, has appeared in the inferences of a robot so poorly informed that it does not even have the concept of repetitions and sampling! In other words, the binomial distribution has an exact *combinatorial* basis, completely independent of the notion of "repetitive sampling".

This gives us a clue toward understanding how the poorly informed robot functions. In conventional probability theory, starting with James Bernoulli (1713), the binomial distribution has always been derived from the postulate that the probability of any result is to be the same at each trial, *strictly independently of what happens at any other trial*. But as we have noted already, that is exactly what the poorly informed robot would say – not out of its knowledge of the physical conditions of the experiment, but out of its complete *ignorance* of what is happening.

Now we could go through many other derivations and we would find that this agreement persists: the poorly informed robot will find not only the binomial but also its generalization, the multinomial distribution, as combinatorial theorems. Then all the usual probability distributions of sampling theory (Poisson, Gamma, Gaussian, Chi–squared *etc.*) will follow as limiting forms of these, as noted in Appendix E. All the results that conventional probability theory has been obtaining from the frequency definition and the assumption of strict independence of different trials, are just what the poorly informed robot would find in the same problem. In other words, we can now characterize the conventional frequentist probability theory functionally, simply as *the reasoning of the poorly informed robot*.

Exercise 9.3: Derive the multinomial distribution found in Chapter 3, Eq. (3–77), as a generalization or extension of our derivation of (9–42).

Then, since the poorly informed robot is unable to do inductive reasoning, we begin to understand why conventional probability theory has trouble with it. Both lack the essential ingredient required for induction; until we learn how to introduce some kind of correlation between the results of different trials, the results of any trials cannot tell us anything about any other trial, and it will be impossible to "take the hint." Indeed, frequentist probability theory is stuck with independent trials because it lays great stress on limit theorems, and examination of them shows that their validity depends entirely on the strict independence of different trials. The slightest positive correlation between the results of different trials will render those theorems qualitatively wrong. Indeed, without that strict independence virtually all of the sampling distributions for estimators, on which orthodox statistics depends, would be incorrect, invalidating their procedures.

Yet on second glance there is an important difference; in conventional probability theory that "independence" is held to mean causal physical independence; to the robot it means logical independence, a very much stronger condition. But from the standpoint of the frequentist, that is only a philosophical difference – not really a functional one – because he confines himself to what we consider conceptually simple problems. We note this particularly in Chapter 16, comparing the work of R. A. Fisher and H. Jeffreys.

Relation to Generating Functions

Note that the number of conceivable outcomes can be written as $N = m^n = Z^n(0)$, so that (9–40) becomes

$$\frac{Z^n(\lambda)}{Z^n(0)} = \sum_{r=0}^{n} b(r|n,p)\, z^r \tag{9–43}$$

where $z \equiv e^{-\lambda}$. This is just what we called the "generating function" for the binomial distribution in Chapter 3 without further explanations.

In any problem we may set $z = e^{-\lambda}$, and instead of a partition function, define a generating function $\Phi(z) \equiv Z(\lambda)/Z(0)$. Of course, anything that can be done with one function can be done also with the other; but in calculations such as (9–23) where one must carry out integrations over complex values of λ or z, the partition function is generally a more convenient tool because it remains single–valued in the complex λ–plane in conditions (*i.e.*, when the g_j are irrational numbers) where the generating function would develop an infinite number of Riemann surfaces in the z–plane.

We have seen above how the partition function may be used to calculate exact results in probability theory. However, its real power appears in problems so complicated that we would not attempt to calculate the exact $Z(\lambda)$ analytically. When n becomes large, there are very accurate asymptotic formulas for $\log Z(\lambda)$ which are amenable to hand calculation. Indeed, partition functions and generating functions are such powerful calculational devices that Laplace's *Théorie analytique des probabilités* devotes Volume 1 entirely to developing the theory of generating functions, and how to use them for solving finite difference equations such as (9–19), before even mentioning probability.

Since the fundamental work of Gibbs (1902), the partition function has also been the standard device on which all useful calculations in Statistical Mechanics are based; indeed, there is hardly any nontrivial problem which can be solved at all without it. Typically, one expresses Z or $\log Z$ as a contour integral, then chooses the path of integration to pass over a saddle point that becomes sharper as $n \to \infty$, whereupon saddle–point integration yields excellent asymptotic formulas. We shall see examples presently.

Then Shannon (1948) found that the difference equation (9–19) and the above way of solving it are the basic tools for calculating channel capacity in Communication Theory. Finally, it is curious that Laplace's original discussion of generating functions contains almost all the mathematical material that Electrical Engineers use today in the theory of digital filters, not thought of as related to probability theory at all.

From Laplace transform theory, the path of integration in (9–23) will be from $(-i\infty)$ to $(i\infty)$ in the complex λ – plane, passing to the right of all singularities in the integrand. In complicated problems one may use the integral representation (9–23) to evaluate probabilities. In particular, integral representations of a function usually provide the easiest way of extracting asymptotic forms (for large n). However, resort to (9–23) is not always necessary if we note the following.

Another Way of Looking at it

The following observation gives us a better intuitive understanding of the partition function method. Unfortunately, it is only a number–theoretic trick, useless in practice. From (9–24) and (9–25) we see that the multiplicity of ways in which the total G can be realized can be written as

$$M(n,G) = \sum_{\{n_j\}} W(n_1 \cdots n_m) \tag{9–44}$$

where we are to sum over all sets of non–negative integers $\{n_j\}$ satisfying

$$\sum n_j = n\ , \qquad \sum n_j g_j = G\ . \tag{9–45}$$

Let $\{n_j\}$ and $\{n'_j\}$ be two such different sets which yield the same total: $\sum n_j g_j = \sum n'_j g_j = G$. Then it follows that

$$\sum_{j=1}^{m} k_j g_j = 0 \tag{9–46}$$

where by hypothesis the integers $k_j \equiv n_j - n'_j$ cannot all be zero.

Two numbers f, g are said to be *incommensurable* if their ratio is not a rational number; *i.e.*, if (f/g) cannot be written as (r/s) where r and s are integers (but of course, any ratio may be thus approximated arbitrarily closely by choosing r, s large enough). Likewise, we shall call the numbers $(g_1 \cdots g_m)$ *jointly incommensurable* if no one of them can be written as a linear combination of the others with rational coefficients. But if this is so, then (9–46) implies that all $k_j = 0$:

$$n_j = n'_j, \qquad 1 \le j \le m$$

so if the $\{g_1 \cdots g_m\}$ are jointly incommensurable, then *in principle* the solution is immediate; for then a given value of $G = \sum n_j g_j$ can be realized by only one set of sample numbers n_j; *i.e.*, if G is specified exactly, this determines the exact values of all the $\{n_j\}$. Then we have only one term in (9–44):

$$M(n,G) = W(n_1 \cdots n_m) \tag{9–47}$$

and

$$M(n-1, G-g_j) = W(n'_1 \cdots n'_m) \tag{9–48}$$

where, necessarily, $n'_i = n_i - \delta_{ij}$. Then the exact result (9–18) reduces to

$$p(r_k = j|G,n,I_0) = \frac{W(n'_1 \cdots n'_m)}{W(n_1 \cdots n_m)} = \frac{(n-1)!}{n!}\,\frac{n_j!}{(n_j-1)!} = \frac{n_j}{n} \tag{9–49}$$

In this case the result could have been found in a different way: whenever by any means the robot knows the sample number n_j (*i.e.*, the number of digits $\{r_1 \cdots r_n\}$ equal to j) but does not know at which trials the j'th result occurred (*i.e.*, which digits are equal to j), it can apply Bernoulli's rule (9–3) directly:

$$P(r_k = j|n_j, I_0) = \frac{n_j}{(total\ number\ of\ digits)} \tag{9–50}$$

Again, the *probability* of any proposition A is equal to the *frequency* with which it is true in the relevant set of equally possible hypotheses. So again our robot, even if poorly informed, is nevertheless producing the standard results that current conventional treatments all assure us are correct. Conventional writers appear to regard this as a kind of law of physics; but we need not invoke any "law" to account for the fact that a measured frequency often approximates an assigned probability (to a relative accuracy something like $n^{-1/2}$ where n is the number of trials). If the information used to assign that probability includes all of the systematic effects at work in the real experiment, then the great majority of all things that *could* happen in the experiment correspond to frequencies remaining in such a shrinking interval; this is simply a combinatorial theorem, which in essence was given already by de Moivre and Laplace in the 18'th Century, in their asymptotic formula. In virtually all of current probability theory this strong connection between probability and frequency is taken for granted for all probabilities but without any explanation of the mechanism that produces it; for us, this connection is only a special case.

Now if certain factors are not varying from one trial to the next, there is presumably some physical cause which is preventing that variation. Therefore, we might call the unvarying factors the *constraints* or the *signal*, the uncontrolled variable factors the *noise* operating in the experiment. Evidently, if we know the constraints in advance, then we can do a tolerably good job of predicting the data. Conversely, given some data we are often interested primarily in estimating what signal is present in them; *i.e.*, what constraints must be operating to produce such data.

The Better Informed Robot

With the clues just uncovered, we are able to educate the robot so that it can do inductive reasoning in more or less the same way that you and I do. Perhaps the best explored, and to date most useful, classes of correlated sampling distributions are those called *Dirichlet*, *exchangeable*, *autoregressive*, and *maximum entropy* distributions. Let us see how each of these enables the robot to deal with problems like the survey sampling noted above.

********************* MUCH MORE COMING HERE! *************************

We can now sum up what we have learned about probability and frequency.

Probability and Frequency

In our terminology, a *probability* is something that we assign, in order to represent a state of knowledge, or that we calculate from previously assigned probabilities according to the rules of probability theory. A *frequency* is a factual property of the real world that we measure or estimate. The phrase "estimating a probability" is just as much a logical incongruity as "assigning a frequency" or "drawing a square circle".

The fundamental, inescapable distinction between probability and frequency lies in this relativity principle: probabilities change when we change our state of knowledge; frequencies do not. It follows that the probability $p(E)$ that we assign to an event E can be equal to its frequency $f(E)$ only for certain particular states of knowledge. Intuitively, one would expect this to be the case when the only information we have about E consists of its observed frequency; and the mathematical rules of probability theory confirm this in the following way.

We note the two most familiar connections between probability and frequency. Under the assumption of exchangeability and certain other prior information (Jaynes, 1968), the rule for translating an observed frequency in a binary experiment into an assigned probability is Laplace's rule of succession. We have encountered this already in Chapter 6 in connection with Urn sampling, and we analyze it in detail in Chapter 18. Under the assumption of independence, the rule for translating an assigned probability into an estimated frequency is Bernoulli's weak law of large numbers (or, to get an error estimate, the de Moivre – Laplace limit theorem).

However, many other connections exist. They are contained, for example, in the principle of maximum entropy (Chapter 11), the principle of transformation groups (Chapter 12), and in the theory of fluctuations in exchangeable sequences (Jaynes, 1978).

If anyone wished to research this matter, we think he could find a dozen logically distinct connections between probability and frequency, that have appeared in various applications. But these connections always appear automatically, as mathematical consequences of probability theory as logic, whenever they are relevant to the problem; there is never any need to define a probability as a frequency.

Indeed, Bayesian theory may justifiably claim to use the notion of frequency more effectively than does the "frequency" theory. For the latter admits only one kind of connection between probability and frequency, and has trouble in cases where a different connection is appropriate. Those cases include some important, real problems which are today at the forefront of new applications.

Today, Bayesian practice has far outrun the original class of problems where frequency definitions were usable; yet it includes as special cases all the useful results that had been found in the frequency theory. In discarding frequency definitions, then, we have not lost "objectivity"; rather, we have advanced to the flexibility of a far deeper kind of objectivity than that envisaged by Venn, von Mises, and Fisher. This flexibility is necessary for scientific inference; for most real problems arise out of incomplete information, and have nothing to do with random experiments.

In physics, when probabilities are allowed to become physically real, logical consistency eventually forces one to regard ordinary objects such as atoms, as unreal; this is rampant in the current literature of statistical mechanics and theoretical physics. In economics, where experiments cannot be repeated, belief that probabilities are real would force one to invent an ensemble of imaginary worlds to define a sample space, diverting attention away from the one real world that we are trying to reason about.

The "propensity" lies not in the definition of probability in general, or in any "physical reality" of probabilities; it lies in the prior information that was used to calculate the probability. Where the appropriate prior information is lacking, so is the propensity. We found already in Chapter 3 that conditional probabilities – even sampling probabilities – express fundamentally *logical inferences* which may or may not correspond to causal physical influences.

****

R. A. Fisher, J. Neyman, R. von Mises, Wm. Feller, and L. J. Savage denied vehemently that probability theory is an extension of logic, and accused Laplace and Jeffreys of committing metaphysical nonsense for thinking that it is. It seems to us that, if Mr. A wishes to study properties of frequencies in random experiments and publish the results for all to see and teach them to the next generation, he has every right to do so, and we wish him every success. But in turn Mr. B has an equal right to study problems of logical inference that have no necessary connection with frequencies or random experiments, and to publish his conclusions and teach them. The world has ample room for both.

Then why should there be such unending conflict, unresolved after over a Century of bitter debate? Why cannot both coexist in peace? What we have never been able to comprehend is this: If Mr. A wants to talk about frequencies, then why can't he just use the *word* "frequency"? Why does he insist on appropriating the word "probability" and using it in a sense that flies in the face of both historical precedent and the common colloquial meaning of that word? By this practice he guarantees that his meaning will be misunderstood by almost every reader who does not belong to his inner circle clique. It seems to us that he would find it easy – and very much in his own self–interest – to avoid these constant misunderstandings, simply by saying what he means. [H. Cramér (1946) did this fairly often, although not with 100% reliability, so his work is today easier to read and comprehend.]

Of course, von Mises, Feller, Fisher, and Neyman would not be in full agreement among themselves on anything. Nevertheless, whenever any of them uses the word "probability", if we merely substitute the word "frequency" we shall go a long way toward clearing up the confusion by producing a statement that means more nearly what they had in mind.

However, we think it is obvious that the vast majority of the real problems of science fall into Mr. B's category and therefore, in the future, science will be obliged to turn more and more toward his viewpoint and results. Furthermore, Mr. B's use of the word "probability" as expressing human information enjoys not only the historical precedent, but it is also closer to the colloquial meaning of the word.

Halley's Mortality Table

An early example of the use of observed frequencies as probabilities, in a more useful and dignified context than gambling, and by a procedure that is so nearly correct that we could not improve on it appreciably today, was provided by the astronomer Edmund Halley (1656–1742) of "Halley's Comet" fame. Interested in many things besides astronomy, he also prepared in 1693 the first modern Mortality Table. Let us dwell a moment on the details of this work because of its great historical interest.

The subject does not quite start with Halley, however. In England, due presumably to increasing population densities, various plagues were rampant from the 16'th Century up to the adoption of public sanitation policies and facilities in the mid 19'th Century. In London, starting intermittently in 1591, and continuously from 1604 for several decades, there were published weekly Bills of Mortality, which listed for each parish the number of births and deaths of males and females and the statistics compiled by the *Searchers*, a body of "antient Matrons" who carried out the unpleasant task of examining corpses and from the physical evidence and any other information they were able to elicit by inquiry, judged as best as they could the cause of each death.

In 1662, John Graunt (1620–1674) called attention to the fact that these Bills, in their totality, contained valuable demographic information that could be useful to Governments and Scholars for many other purposes besides judging the current state of public health.[†] He aggregated the data for 1632 into a single more useful table and made the observation that in sufficiently large pools of data on births there are always slightly more boys than girls, which circumstance provoked many speculations and calculations by probabilists for the next 150 years. Graunt was not a scholar, but a self–educated shopkeeper. Nevertheless, his short work contained so much valuable good sense that it came to the attention of Charles II, who as a reward ordered the Royal Society (which he had founded shortly before) to admit Graunt as a Fellow.[‡]

[†] It appears that this story may be repeated some 330 years later, in the recent realization that the records of credit card companies contain a wealth of economic data which have been sitting there unused for many years. For the largest such company (Citicorp), a record of one percent of the nation's retail sales comes into its computers every day. For predicting some economic trends and activity this is far more detailed, reliable, and timely than the monthly Government releases.

[‡] Contrast this enlightened attitude and behavior with that of Oliver Cromwell shortly before, who through his henchmen did more wanton, malicious damage to Cambridge University than any other person in history. The writer lived for a year in the Second Court of St. John's College, Cambridge, which Cromwell appropriated and put to use, not for scholarly pursuits, but as the stockade for holding his prisoners. Whatever one may think of the private escapades of Charles II, one must ask also: What was the alternative? Had the humorless fanatic Cromwell prevailed, there would have been no Royal Society, and no recognition

Edmund Halley (1656–1742) was highly educated, mathematically competent (later succeeding Wallis (1703) as Savilian Professor of Mathematics at Oxford University and Flamsteed (1720) as Astronomer Royal and Director of the Greenwich Observatory), a personal friend of Isaac Newton and the one who had persuaded him to publish his *Principia* by dropping his own work to see it through publication and paying for it out of his own modest fortune. He was eminently in a position to do more with demographic data than was John Graunt.

In undertaking to determine the actual distribution of age in the population, Halley had extensive data on births and deaths from London and Dublin. But records of the age at death were often missing, and he perceived that London and Dublin were growing rapidly by in–migration, biasing the data with people dying there who were not born there. So he found instead five years' data (1687–1691) for a city with a stable population: Breslau in Silesia (today called Wroclaw, in what is now Poland). Silesians, more meticulous in record keeping and less inclined to migrate, generated better data for his purpose.

Of course, contemporary standards of nutrition, sanitation, and medical care in Breslau might differ from those in England. But in any event Halley produced a mortality table surely valid for Breslau and presumably not badly in error for England. We have converted it into a graph, with three emendations described below, and present it in Fig. 9.1.

In the 17'th Century, even so learned a man as Halley did not have the habits of full, clear expression that we expect in scholarly works today. In reading his work we are exasperated at the ambiguities and omissions, which make it impossible to ascertain some important details about his data and procedure. We know that his data consisted of monthly records of the number of births and deaths and the age of each person at death. Unfortunately, he does not show us the original, unprocessed data, which would today be of far greater value to us than anything in his work, because with modern probability theory and computers, we could easily process the data for ourselves, and extract much more information from them than Halley did.

Halley presents two tables derived from the data, giving respectively the estimated number $d(x)$ of annual deaths (total number /5) at each age of x years (but which inexplicably contains some entries that are not multiples of 1/5), and the estimated distribution $n(x)$ of population by age. Thus the first table is, crudely, something like the negative derivative of the second. But, inexplicably, he omits the very young ($< 7\ yr$) from the first table, and the very old ($> 84\ yr$) from the second, thus withholding what are in many ways the most interesting parts, the regions of strong curvature of the graph.

Even so, if we knew the exact procedure by which he constructed the tables from the raw data, we might be able to reconstruct both tables in their entirety. But he gives absolutely no information about this, saying only, "From these Considerations I have formed the *adjoyned Table*, whose Uses are manifold, and give a more just *Idea* of the *State* and *Condition of Mankind*, than any thing yet extant that I know of." But he fails to inform us what "these Considerations" are, so we are reduced to conjecturing what he actually did.

Although we were unable to find any conjecture which is consistent with all the numerical values in Halley's tables, we can clarify things to some extent. In the first place, the actual number of deaths at each age in the first table naturally shows considerable "statistical fluctuations" from one age to the next. Halley must have done some kind of smoothing of this, because the fluctuations do not show in the second table.

From other evidence in his article we infer that he reasoned as follows: if the population distribution is stable (exactly the same next year as this year), then the difference $n(25) - n(26)$

for scholarly accomplishment in England; quite likely, the magnificent achievements of British science in the 19'th Century would not have happened. It is even problematical whether Cambridge and Oxford Universities would still exist today.

between number now alive at ages 25 and 26 must be equal to the number $d(25)$ now at age 25 who will die in the next year. Thus we would expect that the second table might be constructed by starting with the estimated number (1238) born each year as $n(0)$, and by recursion taking $n(x) = n(x-1) - \overline{d}(x)$, where $\overline{d}(x)$ is the smoothed estimate of d. Finally, the total population of Breslau is estimated as $\sum_x n(x) = 34,000$. But although the later parts of table 2 are well accounted for by this surmise, the early parts $(0 < x < 7)$ do not fit it, and we have been unable to form even a conjecture about how he determined the first six entries of table 2.

Secondly, we have shifted the ages downward by one year in our graph because it appears that the common meanings of terms have changed in 300 years. Today, when we say colloquially that a boy is 'eight years old', we mean that his exact age x is in the range $(8 \leq x < 9)$; *i.e.*, he is actually in his ninth year of life. But we can make sense out of Halley's numbers only if we assume that for him the phrase 'eight years current' meant in the eighth year of life; $(7 < x \leq 8)$.

These points were noted also by Major Greenwood (1942), whose analysis confirms our conclusion about the meaning of 'age current'. However, our attempt to follow his reasoning beyond that point leaves us more confused than before (he suggests that Halley took into account that the death rate of very young children is greater in the first half of a year than in the second; but while we accept the phenomenon, we are unable to see how this could affect his tables, which refer only to whole years). At this point we must give up, and simply accept Halley's judgment, whatever it was.

In Fig. 9.1 we give Halley's second table as a graph of a shifted function $n(y)$. Thus where Halley's table reads (25 567) we give it as $n(24) = 567$, which we interpret to mean an estimated 567 persons in the age range $(24 \leq x < 25)$. Thus our $n(y)$ is what we believe to be Halley's estimated number of persons in the age range $(y, y+1)$ years.

Thirdly, Halley's second table stops at the entry (84 20); yet the first table has data beyond that age, which he used in estimating the total population of Breslau. His first table indicates what we interpret as 19 deaths in the range (85, 100) in the five years, including three at "age current" 100. He estimated the total population in that age range as 107. We have converted this meager information, plus other comparisons of the two tables, into a smoothed extrapolation of Halley's second table [our entries $n(84) \ldots n(99)$], which shows the necessary sharp curvature in the tail.

What strikes us first about this graph is the appalling infant mortality rate. Halley states elsewhere that only 56% of those born survived to the age of six (although this does not agree with his table 2) and that 50% survive to age 17 (which does agree with the table). The second striking feature is the almost perfect linearity in the age range $(35 - 80)$.

Halley notes various uses that can be made of his second table, including estimating the size of the army that the city could raise, and the values of annuities. Let us consider only one, the estimation of future life expectancy. We would think it reasonable to assign a probability that a person of age y will live to age z, as $p = n(z)/n(y)$, to sufficient accuracy.

Actually, Halley does not use the word "probability" but instead refers to "odds" in exactly the same way that we use it today: "- - - if the number of Persons of any *Age* remaining after one year, be divided by the difference between that and the number of the Age proposed, it shews the *odds* that there is, that a Person of that Age does not die in a *Year*." Thus Halley's odds on a person living m more years, given present age of y years is $O(m|y) = n(y+m)/(n(y) - n(y+m)) = p/(1-p)$, in agreement with our calculation.

Another exasperating feature is that Halley pooled the data for males and females, and thus failed to exhibit their different mortality functions; lacking his raw data, we are unable to rectify this.

Let the things which exasperate us in Halley's work be a lesson for us today: the First Commandment of scientific data analysis publication ought to be: "Thou shalt reveal thy full original data, unmutilated by any processing whatsoever." Just as today we could do more with Halley's raw data than he did, future readers may be able to do more with our raw data than we can, if only we will refrain from mutilating it according to our present purposes and prejudices. At the very least, they will approach our data with a different state of prior knowledge than ours, and we have seen how much this can affect the conclusions.

Exercise 9.3. Suppose you had the same raw data as Halley. How would you process them today, taking full advantage of probability theory? How different would the actual conclusions be?

COMMENTS

The Irrationalists. Philosophers have argued over the nature of induction for centuries. Some, from David Hume (1711–1776) in the mid–18'th Century to Karl Popper in the mid–20'th, [for example, Popper & Miller (1983)], have tried to deny the possibility of induction, although all scientific knowledge has been obtained by induction. D. Stove (1982) calls them "the irrationalists" and tries to understand (1) How could such an absurd view ever have arisen? and (2) By what linguistic practices do the irrationalists succeed in gaining an audience? However, since we are not convinced that much of an audience exists, we were concerned above not with exposing the already obvious fallacy of irrationalism, but with showing how probability theory as logic supplies a constructive alternative to it.

In denying the possibility of induction, Popper holds that theories can never attain a high probability. But this presupposes that the theory is being tested against an infinite number of alternatives. We would observe that the number of atoms in the known universe is finite; so also, therefore, is the amount of paper and ink available to write alternative theories. It is not the absolute status of an hypothesis embedded in the universe of all conceivable theories, but the plausibility of an hypothesis *relative to a definite set of specified alternatives*, that Bayesian inference determines.

As we showed in connection with multiple hypothesis testing in Chapter 4, and Newton's theory in Chapter 5, an hypothesis can attain a very high or very low probability *within a class of well–defined alternatives*. Its probability within the class of all conceivable theories is neither large nor small; it is simply undefined because the class of all conceivable theories is undefined. In other words, Bayesian inference deals with determinate problems – not the undefined ones of Popper – and we would not have it otherwise.

The objection to induction is often stated in different terms. If a theory cannot attain a high absolute probability against all alternatives, then there is no way to prove that induction from it will be right. But that quite misses the point; it is not the function of induction to be 'right', and working scientists do not use it for that purpose (and could not if we wanted to). The functional use of induction in science is not to tell us what predictions must be true, but rather *what predictions are most strongly indicated by our present hypotheses and our present information*?

Put more carefully, What predictions are most strongly indicated by the information *that we have put into the calculation*? It is quite legitimate to do induction based on hypotheses that we do not believe; or even that we know to be false, to learn what their predictable consequences would be. Indeed, an experimenter seeking evidence for his favorite theory, does not know what to look for unless he knows what predictions are made by some alternative theory. He must give temporary lip–service to the alternative to find out what it predicts, although he does not really believe it.

If predictions made by a theory are borne out by future observation, then we become more confident of the hypotheses that led to them; and if the predictions never fail in vast numbers of tests, we come eventually to call those hypotheses "physical laws". Successful induction is, of course, of great practical value in planning strategies for the future. But from successful induction we do not learn anything basically new; we only become more confident of what we knew already.

On the other hand, if the predictions prove to be wrong, then induction has served its real purpose; we have learned that our hypotheses are wrong or incomplete, and from the nature of the error we have a clue as to how they might be improved. So those who criticize induction on the grounds that it might not be right, could not possibly be more mistaken. Induction is most valuable to a scientist just when it turns out to be wrong. But to comprehend this, one must recognize that probability distributions do not describe reality; they describe only *our present information about reality* – which is, after all, the only thing we have to reason on.

Some striking case histories are found in biology, where causal relations are often so complex and subtle that it is remarkable that it was possible to uncover them at all. For example, it became clear in the 20'th Century that new influenza pandemics were coming out of China; the worst ones acquired names like the Asian Flu (1957), the Hong Kong Flu (1968), and Beijing A (1993). It appears that the cause has been traced to the fact that Chinese farmers raise ducks and pigs side by side. Humans are not infected directly by viruses in ducks, even by handling them and eating them; but pigs can absorb duck viruses, transfer some of their genes to other viruses, and in this form pass them on to humans, where they take on a life of their own because they appear as something entirely new, for which the human immune system is unprepared.

An equally remarkable causal chain is in the role of the gooseberry as a host transmuting and transmitting the white pine blister rust disease. Many other examples of unravelling subtle cause–effect chains are found in the classic work of Louis Pasteur, and of modern medical researchers who continue to succeed in locating the specific genes responsible for various disorders.

We stress that all of these triumphant examples of highly important detective work were accomplished by qualitative plausible reasoning using the format defined by Pólya (1954). Modern Bayesian analysis is just the unique quantitative expression of this reasoning format; the inductive reasoning that philosophers like Hume and Popper held to be impossible. It is true that this reasoning format does not guarantee that the conclusion *must* be correct; rather, it tells us which conclusions are indicated most strongly *by our present information*, whereupon direct tests can confirm it or refute it. Without the preparatory inductive reasoning phase, one would not know which direct tests to try.

Superstitions

Another curious circumstance is that, although induction has proved a tricky thing to understand and justify logically, the human mind has a predilection for rampant, uncontrolled induction, and it requires much education to overcome this. As we noted briefly in Chapter 5, the reasoning of those without training in any mental discipline – who are therefore unfamiliar with either deductive logic or probability theory – is mostly unjustified induction.

In spite of modern science, general human comprehension of the world has progressed very little beyond the level of ancient superstitions. As we observe constantly in news commentaries and documentaries, the untrained mind never hesitates to interpret every observed correlation as a causal influence, and to predict its recurrence in the future. For one with no comprehension of what science is, it makes no difference whether that causation is or is not explainable rationally by a physical mechanism. Indeed, the very idea that a causal influence requires a physical mechanism to bring it about, is quite foreign to the thinking of the uneducated; belief in supernatural influences makes such hypotheses, for them, unnecessary.[†]

[†] In the meantime, progress in human knowledge continues to be made by those who, like modern biologists, do think in terms of physical mechanisms; as soon as that premise is abandoned, progress ceases, as we observe in modern quantum theory.

Thus the commentators for the very numerous TV Nature documentaries showing us the behavior of animals in the wild, never hesitate to see in every random mutation some teleological purpose; always, the environmental niche is there and the animal mutates, purposefully, in order to adapt to it. Each conformation of feather, beak, and claw is explained to us in terms of its *purpose*, but never suggesting how an unsubstantial purpose could bring about a physical change in the animal.[‡]

It would seem that we have here a golden opportunity to illustrate and explain evolution; yet the commentators have no comprehension of the simple, easily understood cause–and–effect mechanism pointed out by Charles Darwin. When we have the palpable evidence, and a simple explanation of it, before us, it is incredible that anybody could look to something supernatural, that nobody has ever observed, to explain it. But never does a commentator imagine that the mutation occurs first, and the resulting animal is obliged to seek a niche where it can survive and use its body structures as best it can in that environment. We see only the ones who were successful at this; the others are not around when the cameraman arrives and their small numbers make it highly unlikely that a paleontologist will ever find evidence of them.[*] These documentaries always have very beautiful photography, and they deserve commentaries that make sense.

Indeed, there are powerful counter–examples to the theory that an animal adapts its body structure purposefully to its environment. In the Andes mountains there are woodpeckers where there are no trees. Evidently, they did not become woodpeckers by adapting their body structures to their environment; rather, they were woodpeckers first who, finding themselves through some accident in a strange environment, survived by putting their body structures to a different use. In our view, this is not an exceptional case; rather it is a common feature of almost all evolution.

[‡] But it is hard to believe that the ridiculous color patterns of the Puffin, the Wood Duck, and the Pileated Woodpecker serve any survival purpose; what would the teleologists have to say about this? Our answer would be that, even without subsequent natural selection, divergent evolution can proceed by mutations that have nothing to do with survival. We noted some of this in Chapter 7, in connection with the work of Francis Galton.

[*] But a striking exception was found in the Burgess shale of the Canadian Rockies (Gould, 1989), in which beautifully preserved fossils of soft–bodied creatures contemporary with trilobites, which did not survive to leave any evolutionary lines, were found in such profusion that it radically revised our picture of life in the Cambrian.

cc10k, 5/11/96

CHAPTER 10

PHYSICS OF “RANDOM EXPERIMENTS”

> *“I believe, for instance, that it would be very difficult to persuade an intelligent physicist that current statistical practice was sensible, but that there would be much less difficulty with an approach via likelihood and Bayes' theorem.”* — G. E. P. Box (1962).

As we have noted several times, the idea that probabilities are physically real things, based ultimately on observed frequencies of random variables, underlies most recent expositions of probability theory, which would seem to make it a branch of experimental science. At the end of Chapter 8 we saw some of the difficulties that this view leads us to; in some real physical experiments the distinction between random and nonrandom quantities is so obscure and artificial that you have to resort to black magic in order to force this distinction into the problem at all. But that discussion did not reach into the serious physics of the situation. In this Chapter, we take time off for an interlude of physical considerations that show the fundamental difficulty with the notion of “random” experiments.

An Interesting Correlation

There have always been dissenters from the “frequentist” view who have maintained, with Laplace, that probability theory is properly regarded as the “calculus of inductive reasoning,” and is not fundamentally related to random experiments at all. A major purpose of the present work is to demonstrate that probability theory can deal, consistently and usefully, with far more than frequencies in random experiments, if only it is allowed to do so. According to this view, consideration of random experiments is only one specialized *application* of probability theory, and not even the most important one; for probability theory as logic solves far more general problems of reasoning which have nothing to do with chance or randomness, but a great deal to do with the real world. In the present Chapter we carry this further and show that ‘frequentist’ probability theory has major logical difficulties in dealing with the very random experiments for which it was invented.

One who studies the literature of these matters perceives that there is a strong correlation; those who have advocated the non–frequency view have tended to be physicists, while up until very recently mathematicians, statisticians, and philosophers almost invariably favored the frequentist view. Thus it appears that the issue is not merely one of philosophy or mathematics; in some way not yet clear, it also involves physics.

The mathematician tends to think of a random experiment as an abstraction – really nothing more than a sequence of numbers. To define the “nature” of the random experiment he introduces statements – variously termed assumptions, postulates, or axioms – which specify the sample space and assert the existence, and certain other properties, of limiting frequencies. But in the real world, a random experiment is not an abstraction whose properties can be defined at will. It is surely subject to the laws of physics; yet recognition of this is conspicuously missing from frequentist expositions of probability theory. Even the phrase ‘laws of physics’ is not to be found in them. But defining a probability as a frequency is not merely an excuse for ignoring the laws of physics; it is more serious than that. We want to show that maintenance of a frequency interpretation to the exclusion of all others *requires* one to ignore virtually all the professional knowledge that scientists have about real phenomena. If the aim is to draw inferences about real phenomena, this is hardly the way to begin.

As soon as a specific random experiment is described, it is the nature of a physicist to start thinking, not about the abstract sample space thus defined, but about the physical mechanism of the phenomenon being observed. The question whether the usual postulates of probability theory are compatible with the known laws of physics is capable of logical analysis, with results that have a direct bearing on the question, not of the mathematical consistency of frequency and non–frequency theories of probability, but of their applicability in real situations. In our opening quotation, the statistician G. E. P. Box noted this; let us analyze his statement in the light both of history and of physics.

Historical Background

As we know, probability theory started in consideration of gambling devices by Gerolamo Cardano in the 16'th Century, and by Pascal and Fermat in the 17'th; but its development beyond that level, in the 18'th and 19'th centuries, was stimulated by applications in astronomy and physics, and was the work of people – James and Daniel Bernoulli, Laplace, Poisson, Legendre, Gauss, Boltzmann, Maxwell, Gibbs – most of whom we would describe today as mathematical physicists.

But reactions against Laplace started already in the mid Nineteenth Century, when Cournot, Ellis, Boole, and Venn – none of whom had any training in physics – were unable to comprehend Laplace's rationale and attacked what he did, simply ignoring all his successful results. In particular, John Venn, a philosopher without the tiniest fraction of Laplace's knowledge of either physics or mathematics, nevertheless considered himself competent to write scathing, sarcastic attacks on Laplace's work. In Chapter 16 we note his possible later influence on the young R. A. Fisher. Boole (1854, Chapters XX and XXI) shows repeatedly that he does not understand the function of Laplace's prior probabilities (to represent a *state of knowledge* rather than a physical fact). In other words, he too suffers from the Mind Projection Fallacy. On p. 380 he rejects a uniform prior probability assignment as 'arbitrary' and explicitly *refuses* to examine its consequences; by which tactics he prevents himself from learning what Laplace was really doing and why.

Laplace was defended staunchly by the mathematician Augustus de Morgan and the physicist W. Stanley Jevons,† who understood Laplace's motivations and for whom his beautiful mathematics was a delight rather than a pain. Nevertheless, the attacks of Boole and Venn found a sympathetic hearing in England among non–physicists. Perhaps this was because biologists, whose training in physics and mathematics was for the most part not much better than Venn's, were trying to find empirical evidence for Darwin's theory and realized that it would be necessary to collect and analyze large masses of data in order to detect the small, slow trends that they visualized as the means by which evolution proceeds. Finding Laplace's mathematical works too much to digest, and since the profession of Statistician did not yet exist, they would naturally welcome suggestions that they need not read Laplace after all.

In any event, a radical change took place at about the beginning of this Century when a new group of workers, not physicists, entered the field. They were concerned mostly with biological problems and with Venn's encouragement proceeded to reject virtually everything done by Laplace. To fill the vacuum, they sought to develop the field anew based on entirely different principles in which one assigned probabilities only to data and to nothing else. Indeed, this did simplify the mathematics at first, because many of the problems solvable by Laplace's methods now lay outside the ambit of their methods. As long as they considered only relatively simple problems (technically, problems with sufficient statistics but no nuisance parameters), the shortcoming was

† Jevons did so many things that it is difficult to classify him by occupation. Zabell (1989), apparently guided by the title of one of his books (1874), describes Jevons as a logician and philosopher of science; from examination of his other works we are inclined to list him rather as a physicist who wrote extensively on economics.

not troublesome. This extremely aggressive school soon dominated the field so completely that its methods have come to be known as "orthodox" statistics, and the modern profession of Statistician has evolved mostly out of this movement.

Simultaneously with this development, the physicists – with Sir Harold Jeffreys as almost the sole exception – quietly retired from the field, and statistical analysis disappeared from the physics curriculum. This disappearance has been so complete that, if today someone were to take a poll of physicists, we think that not one in a hundred could identify such names as Fisher, Neyman, Wald; or such terms as maximum likelihood, confidence interval, analysis of variance.

This course of events – the leading role of physicists in development of the original Bayesian methods, and their later withdrawal from orthodox statistics – was no accident. As further evidence that there is some kind of basic conflict between orthodox statistical doctrine and physics, we may note that two of the most eloquent proponents of non–frequency definitions in the early 20'th Century – Poincaré and Jeffreys – were mathematical physicists of the very highest competence, as was Laplace. Professor Box's statement thus has a clear basis in historical fact.

But what is the nature of this conflict? What is there in the physicist's knowledge that leads him to reject the very thing that the others regard as conferring "objectivity" on probability theory? To see where the difficulty lies, we examine a few simple random experiments from the physicist's viewpoint. The facts we want to point out are so elementary that one cannot believe they are really unknown to modern writers on probability theory. The continual appearance of new textbooks which ignore them merely illustrates what we physics teachers have always known; you can teach a student the laws of physics, but you cannot teach him the art of recognizing the *relevance* of this knowledge, much less the habit of actually *applying* it, in his everyday problems.

How to Cheat at Coin and Die Tossing

Cramér (1946) takes it as an axiom that "Any random variable has a unique probability distribution." From the later context, it is clear that what he really means is that it has a unique *frequency* distribution. If one assumes that the number obtained by tossing a die is a random variable, this leads to the conclusion that the frequency with which a certain face comes up is a physical property of the die; just as much so as its mass, moment of inertia, or chemical composition. Thus, Cramér (p. 154) states:

> "The numbers p_r should, in fact, be regarded as physical constants of the particular die that we are using, and the question as to their numerical values cannot be answered by the axioms of probability theory, any more than the size and the weight of the die are determined by the geometrical and mechanical axioms. However, experience shows that in a well–made die the frequency of any event r in a long series of throws usually approaches 1/6, and accordingly we shall often assume that all the p_r are equal to 1/6 $\cdots$."

To a physicist, this statement seems to show utter contempt for the known laws of mechanics. The results of tossing a die many times do *not* tell us any definite number characteristic only of the die. They tell us also something about how the die was tossed. If you toss "loaded" dice in different ways, you can easily alter the relative frequencies of the faces. With only slightly more difficulty, you can still do this if your dice are perfectly "honest."

Although the principles will be just the same, it will be simpler to discuss a random experiment with only two possible outcomes per trial. Consider, therefore, a "biased" coin, about which I. J. Good (1962) has remarked:

> "Most of us probably think about a biased coin as if it had a physical probability. Now whether it is defined in terms of frequency or just falls out of another type of theory, I think we do argue that way. I suspect that even the most extreme subjectivist such as de Finetti would have to agree that he did sometimes think that way, though he would perhaps avoid doing it in print."

We do not know de Finetti's private thoughts, but would observe that it is just the famous exchangeability theorem of de Finetti which shows us how to carry out a probability analysis of the biased coin *without* thinking in the manner suggested.

In any event, it is easy to show how a physicist would analyze the problem. Let us suppose that the center of gravity of this coin lies on its axis, but displaced a distance x from its geometrical center. If we agree that the result of tossing this coin is a "random variable," then according to the axiom stated by Cramér and hinted at by Good, there must exist a definite functional relationship between the frequency of heads and x:

$$p_H = f(x) \, . \tag{10–1}$$

But this assertion goes far beyond the mathematician's traditional range of freedom to invent arbitrary axioms, and encroaches on the domain of physics; for the laws of mechanics are quite competent to tell us whether such a functional relationship does or does not exist.

The easiest game to analyze turns out to be just the one most often played to decide such practical matters as the starting side in a football game. Your opponent first calls "heads" or "tails" at will. You then toss the coin into the air, catch it in your hand, and without looking at it, show it first to your opponent, who wins if he has called correctly. It is further agreed that a "fair" toss is one in which the coin rises at least nine feet into the air, and thus spends at least 1.5 seconds in free flight.

The laws of mechanics now tell us the following. The ellipsoid of inertia of a thin disc is an oblate spheroid of eccentricity $1/\sqrt{2}$. The displacement x does not affect the symmetry of this ellipsoid, and so according to the Poinsot construction, as found in textbooks on rigid dynamics [such as Routh (1955) or Goldstein (1980, Chapter 5)], the polhodes remain circles concentric with the axis of the coin. In consequence, the character of the tumbling motion of the coin while in flight is exactly the same for a biased as an unbiased coin, except that for the biased one it is the center of gravity, rather than the geometrical center, which describes the parabolic "free particle" trajectory.

An important feature of this tumbling motion is conservation of angular momentum; during its flight the angular momentum of the coin maintains a fixed direction in space (but the angular *velocity* does not; and so the tumbling may appear chaotic to the eye). Let us denote this fixed direction by the unit vector n; it can be any direction you choose, and it is determined by the particular kind of twist you give the coin at the instant of launching. Whether the coin is biased or not, it will show the same face throughout the motion if viewed from this direction (unless, of course, n is exactly perpendicular to the axis of the coin, in which case it shows no face at all).

Therefore, in order to know which face will be uppermost in your hand, you have only to carry out the following procedure. Denote by k a unit vector passing through the coin along its axis, with its point on the "heads" side. Now toss the coin with a twist so that k and n make an acute angle, then catch it with your palm held flat, in a plane normal to n. On successive tosses, you can let the direction of n, the magnitude of the angular momentum, and the angle between n and k, vary widely; the tumbling motion will then appear entirely different to the eye on different tosses, and it would require almost superhuman powers of observation to discover your strategy.

Thus, anyone familiar with the law of conservation of angular momentum can, after some practice, cheat at the usual coin–toss game and call his shots with 100 per cent accuracy. You can obtain any frequency of heads you want; and *the bias of the coin has no influence at all on the results*!

Of course, as soon as this secret is out, someone will object that the experiment analyzed is too "simple." In other words, those who have postulated a physical probability for the biased coin have, without stating so, really had in mind a more complicated experiment in which some kind of "randomness" has more opportunity to make itself felt.

While accepting this criticism, we cannot suppress the obvious comment: scanning the literature of probability theory, isn't it curious that so many mathematicians, usually far more careful than physicists to list all the qualifications needed to make a statement correct, should have failed to see the need for any qualifications here? However, to be more constructive, we can just as well analyze a more complicated experiment.

Suppose that now, instead of catching the coin in our hands, we toss it onto a table, and let it spin and bounce in various ways until it comes to rest. Is this experiment sufficiently "random" so that the true "physical probability" will manifest itself? No doubt, the answer will be that it is not sufficiently random if the coin is merely tossed up six inches starting at the table level, but it will become a "fair" experiment if we toss it up higher.

Exactly how high, then, must we toss it before the true physical probability can be measured? This is not an easy question to answer, and we make no attempt to answer it here. It would appear, however, that anyone who asserts the existence of a physical probability for the coin ought to be prepared to answer it; otherwise it is hard to see what content the assertion has (that is, there is no way to confirm it or disprove it).

We do not deny that the bias of the coin will now have some influence on the frequency of heads; we claim only that the amount of that influence depends very much on how you toss the coin so that, again in this experiment, there is no definite number $p_H = f(x)$ describing a physical property of the coin. Indeed, even the direction of this influence can be reversed by different methods of tossing, as follows.

However high we toss the coin, we still have the law of conservation of angular momentum; and so we can toss it by *Method A*: to ensure that heads will be uppermost when the coin first strikes the table, we have only to hold it heads up, and toss it so that the total angular momentum is directed vertically. Again, we can vary the magnitude of the angular momentum, and the angle between n and k, so that the motion appears quite different to the eye on different tosses, and it would require very close observation to notice that heads remains uppermost throughout the free flight. Although what happens after the coin strikes the table is complicated, the fact that heads is uppermost at first has a strong influence on the result, which is more pronounced for large angular momentum.

Many people have developed the knack of tossing a coin by *Method B*: it goes through a phase of standing on edge and spinning rapidly about a vertical axis, before finally falling to one side or the other. If you toss the coin this way, the eccentric position of the center of gravity will have a dominating influence, and render it practically certain that it will fall always showing the same face. Ordinarily, one would suppose that the coin prefers to fall in the position which gives it the lowest center of gravity; *i.e.,* if the center of gravity is displaced toward tails, then the coin should have a tendency to show heads. However, for an interesting mechanical reason, which we leave for you to work out from the principles of rigid dynamics, method B produces the opposite influence, the coin strongly preferring to fall so that its center of gravity is high.

On the other hand, the bias of the coin has a rather small influence in the opposite direction if we toss it by *Method C*: the coin rotates about a horizontal axis which is perpendicular to the axis of the coin, and so bounces until it can no longer turn over.

In this experiment also, therefore, a person familiar with the laws of mechanics can toss a biased coin so that it will produce predominantly either heads or tails, at will. Furthermore, the effect of method A persists whether the coin is biased or not; and so one can even do this with a perfectly "honest" coin. Finally, although we have been considering only coins, essentially the same mechanical considerations (with more complicated details) apply to the tossing of any other object, such as a die.

The writer has never thought of a biased coin 'as if it had a physical probability' because, being a professional physicist, I know that it does *not* have a physical probability. From the fact that we have seen a strong preponderance of heads, we cannot conclude legitimately that the coin is biased; it may be biased, or it may have been tossed in a way that systematically favors heads. Likewise, from the fact that we have seen equal numbers of heads and tails, we cannot conclude legitimately that the coin is "honest." It may be honest, or it may have been tossed in a way that nullifies the effect of its bias.

Experimental Evidence

Since the conclusions just stated are in direct contradiction to what is postulated, almost universally, in expositions of probability theory, it is worth noting that you can verify them easily in a few minutes of experimentation in your kitchen. An excellent "biased coin" is provided by the metal lid of a small pickle jar, of the type which is not knurled on the outside, and has the edge rolled inward rather than outward, so that the outside surface is accurately round and smooth, and so symmetrical that on an edge view one cannot tell which is the top side.

Suspecting that many people not trained in physics, simply would not believe the things just claimed without experimental proof, we have performed these experiments with a jar lid of diameter $d = 2\ 5/8$ inches, height $h = 3/8$ inch. Assuming a uniform thickness for the metal, the center of gravity should be displaced from the geometrical center by a distance $x = dh/(2d + 8h) = 0.120$ inches; and this was confirmed by hanging the lid by its edge and measuring the angle at which it comes to rest. Ordinarily, one expects this bias to make the lid prefer to fall bottom side (*i.e.,* the inside) up; and so this side will be called "heads." The lid was tossed up about 6 feet, and fell onto a smooth linoleum floor. I allowed myself ten practice tosses by each of the three methods described, and then recorded the results of a number of tosses by: method A deliberately favoring heads, method A deliberately favoring tails, method B, and method C, as given in Table 10.1.

Method	*No. of tosses*	*No. of heads*
A(H)	100	99
A(T)	50	0
B	100	0
C	100	54

Table 10.1. Results of tossing a "biased coin" in four different ways.

In method A the mode of tossing completely dominated the result (the effect of bias would, presumably, have been greater if the "coin" were tossed onto a surface with a greater coefficient of friction). In method B, the bias completely dominated the result (in about thirty of these tosses it looked for a while as if the result were going to be heads, as one might naively expect; but each time the "coin" eventually righted itself and turned over, as predicted by the laws of rigid dynamics). In method C, there was no significant evidence for any effect of bias. The conclusions are pretty clear.

A holdout can always claim that tossing the coin in any of the four specific ways described is "cheating," and that there exists a "fair" way of tossing it, such that the "true" physical probabilities of the coin will emerge from the experiment. But again, the person who asserts this should be prepared to define precisely what this fair method is, otherwise the assertion is without content. Presumably, a fair method of tossing ought to be some kind of random mixture of methods A(H), A(T), B, C, and others; but what is a "fair" relative weighting to give them? It is difficult to see how one could define a "fair" method of tossing except by the condition that it should result in a certain frequency of heads; and so we are involved in a circular argument.

This analysis can be carried much further, as we shall do below; but perhaps it is sufficiently clear already that analysis of coin and die tossing is not a problem of abstract statistics, in which one is free to introduce postulates about “physical probabilities” which ignore the laws of physics. It is a problem of mechanics, highly complicated and irrelevant to probability theory except insofar as it forces us to think a little more carefully about how probability theory must be formulated if it is to be applicable to real situations. Performing a random experiment with a coin does not tell us what the physical probability of heads is; it may tell us something about the bias, but it also tells us something about how the coin is being tossed. Indeed, unless we know how it is being tossed, we cannot draw any reliable inferences about its bias from the experiment.

It may not, however, be clear from the above that conclusions of this type hold quite generally for random experiments, and in no way depend on the particular mechanical properties of coins and dice. In order to illustrate this, consider an entirely different kind of random experiment, as a physicist views it.

Bridge Hands

Elsewhere we quote Professor Wm. Feller’s pronouncements on the use of Bayes’ theorem in quality control testing (Chap.17), on Laplace’s rule of succession (Chap. 18), and on Daniel Bernoulli’s conception of the utility function for decision theory (Chap. 13). He does not fail us here either; in this interesting textbook (Feller, 1951), he writes: “*The number of possible distributions of cards in bridge is almost* 10^{30}. *Usually, we agree to consider them as equally probable. For a check of this convention more than* 10^{30} *experiments would be required—a billion of billion of years if every living person played one game every second, day and night.*” Here again, we have the view that bridge hands possess “physical probabilities,” that the uniform probability assignment is a “convention,” and that the ultimate criterion for its correctness must be observed frequencies in a random experiment.

The thing which is wrong here is that none of us – not even Feller – would be willing to use this criterion with a real deck of cards. Because, if we know that the deck is an honest one, our common sense tells us something which carries more weight than 10^{30} random experiments do. We would, in fact, be willing to accept the result of the random experiment *only if it agreed with our preconceived notion that all distributions are equally likely.*

To many, this last statement will seem like pure blasphemy – it stands in violent contradiction to what we have all been taught is the correct attitude toward probability theory. Yet in order to see why it is true, we have only to imagine that those 10^{30} experiments *had* been performed, and the uniform distribution was not forthcoming. If all distributions of cards have equal frequencies, then any combination of two specified cards will appear together in a given hand, on the average, once in $(52 \times 51)/(13 \times 12) = 17$ deals. But suppose that the combination (Jack of hearts – Seven of clubs) appeared together in each hand three times as often as this. Would we then accept it as an established fact that there is something about the particular combination (Jack of hearts – Seven of clubs) that makes it inherently more likely than others?

We would not. We would reject the experiment and say that the cards had not been properly shuffled. But once again we are involved in a circular argument, because there is no way to define a “proper” method of shuffling except by the condition that it should produce all distributions with equal frequency!

But any attempt to find such a definition involves one in even deeper logical difficulties; one dare not describe the procedure of shuffling in exact detail because that would destroy the “randomness” and make the exact outcome predictable and always the same. In order to keep the experiment “random”, one must describe the procedure incompletely, so that the outcome will be different on different runs. But how could one prove that an incompletely defined procedure will produce all distributions with equal frequency? It seems to us that the attempt to uphold Feller's postulate of physical probabilities for bridge hands leads one into an outright logical contradiction.

Conventional teaching holds that probability assignments must be based fundamentally on frequencies; and that any other basis is at best suspect, at worst irrational with disastrous consequences. On the contrary, this example shows very clearly that *there is a principle for determining probability assignments which has nothing to do with frequencies, yet is so compelling that it takes precedence over any amount of frequency data.* If present teaching does not admit the existence of this principle, it is only because our intuition has run so far ahead of logical analysis – just as it does in elementary geometry – that we have never taken the trouble to present that logical analysis in a mathematically respectable form. But if we learn how to do this, we may expect to find that the mathematical formulation can be applied to a much wider class of problems, where our intuition alone would hardly suffice.

In carrying out a probability analysis of bridge hands, are we really concerned with physical probabilities; or with inductive reasoning? To help answer this, consider the following scenario. The date is 1956, when the writer met Willy Feller and had a discussion with him about these matters. Suppose I had told him that I have dealt at bridge 1000 times, shuffling "fairly" each time; and that in every case the seven of clubs was in my own hand. What would his reaction be? He would, I think, mentally visualize the number

$$\left(\frac{1}{4}\right)^{1000} = 10^{-602} \qquad (10\text{–}2)$$

and conclude instantly that I have not told the truth; and no amount of persuasion on my part would shake that judgment. But what accounts for the strength of his belief? Obviously, it cannot be justified if our assignment of equal probabilities to all distributions of cards (therefore probability 1/4 for the seven of clubs to be in the dealer's hand) is merely a "convention," subject to change in the light of experimental evidence; he rejects my reported experimental evidence, just as we did above. Even more obviously, he is not making use of any knowledge about the outcome of an experiment involving 10^{30} bridge hands.

Then *what is the extra evidence he has*, which his common sense tells him carries more weight than do any number of random experiments; but whose help he refuses to acknowledge in writing textbooks? In order to maintain the claim that probability theory is an experimental science, based fundamentally not on logical inference but on frequency in a random experiment, it is necessary to suppress some of the information which is available. This suppressed information, however, is just what enables our inferences to approach the certainty of deductive reasoning in this example and many others.

The suppressed evidence is, of course, simply our recognition of the *symmetry* of the situation. The only difference between a seven and an eight is that there is a different number printed on the face of the card. Our common sense tells us that where a card goes in shuffling depends only on the mechanical forces that are applied to it; and not on which number is printed on its face. If we observe any systematic tendency for one card to appear in the dealer's hand, which persists on indefinite repetitions of the experiment, we can conclude from this only that there is some systematic tendency in the procedure of shuffling, which alone determines the outcome of the experiment.

Once again, therefore, performing the experiment tells you nothing about the "physical probabilities" of different bridge hands. It tells you something about how the cards are being shuffled. But the full power of symmetry as cogent evidence has not yet been revealed in this argument; we return to it presently.

General Random Experiments

In the face of all the foregoing arguments, one can still take the following position (as a member of the audience did after one of the writer's lectures): "You have shown only that coins, dice, and cards represent exceptional cases, where physical considerations obviate the usual probability postulates; *i.e.,* they are not really 'random experiments.' But that is of no importance because these devices are used only for illustrative purposes; in the more dignified random experiments which merit the serious attention of the scientist, there *is* a physical probability."

To answer this we note two points. First, we reiterate that when anyone asserts the existence of a physical probability in any experiment, then the onus is on him to define the exact circumstances in which this physical probability can be measured; otherwise the assertion is without content.

This point needs to be stressed: those who assert the existence of physical probabilities do so in the belief that this establishes for their position an 'objectivity' that those who speak only of a 'state of knowledge' lack. Yet to assert as fact something which cannot be either proved or disproved by observation of facts, is the opposite of objectivity; it is to assert something that one could not possibly know to be true. Such an assertion is not even entitled to be called a description of a 'state of knowledge'.

Secondly, note that any specific experiment for which the existence of a physical probability is asserted, is subject to physical analysis like the ones just given, which will lead eventually to an understanding of its mechanism. But as soon as this understanding is reached, then this new experiment will also appear as an exceptional case like the above ones, where physical considerations obviate the usual postulates of physical probabilities.

For, as soon as we have understood the mechanism of any experiment E, then there is logically no room for any postulate that various outcomes possess physical probabilities; for the question: "What are the probabilities of various outcomes $(O_1, O_2 \cdots)$?" then reduces immediately to the question: "What are the probabilities of the corresponding initial conditions $(I_1, I_2 \cdots)$ that lead to these outcomes?"

We might suppose that the possible initial conditions $\{I_k\}$ of experiment E themselves possess physical probabilities. But then we are considering an antecedent random experiment E', which produces conditions I_k as its possible outcomes: $I_k = O'_k$. We can analyze the physical mechanism of E' and as soon as this is understood, the question will revert to: "What are the probabilities of the various initial conditions I'_k for experiment E'?"

Evidently, we are involved in an infinite regress $\{E,\ E',\ E'', \cdots\}$; the attempt to introduce a physical probability will be frustrated at every level where our knowledge of physical law permits us to analyze the mechanism involved. The notion of "physical probability" must retreat continually from one level to the next, as knowledge advances.

We are, therefore, in a situation very much like the "warfare between science and theology" of earlier times. For several centuries, theologians with no factual knowledge of astronomy, physics, biology, and geology, nevertheless considered themselves competent to make dogmatic factual assertions which encroached on the domains of those fields – which they were later forced to retract one by one in the face of advancing knowledge.

Clearly, probability theory ought to be formulated in a way that avoids factual assertions properly belonging to other fields, and which will later need to be retracted (as is now the case for many assertions in the literature concerning coins, dice, and cards). It appears to us that the only formulation which accomplishes this, and at the same time has the analytical power to deal with the current problems of science, is the one which was seen and expounded on intuitive grounds by Laplace and Jeffreys. Its validity is a question of logic, and does not depend on any physical assumptions.

As we saw in Chapter 2, a major contribution to that logic was made by R. T. Cox (1946), (1961), who showed that those intuitive grounds can be replaced by theorems. We think it is no accident that Richard Cox was also a physicist (Professor of Physics and Dean of the Graduate School at Johns Hopkins University), to whom the things we have pointed out here would be evident from the start.

The Laplace–Jeffreys–Cox formulation of probability theory does not require us to take one reluctant step after another down that infinite regress; it recognizes that anything which – like the child's spook – continually recedes from the light of detailed inspection, can exist only in our imagination. Those who believe most strongly in physical probabilities, like those who believe in astrology, never seem to ask what would constitute a controlled experiment capable of confirming or disproving their belief.

Indeed, the examples of coins and cards should persuade us that such controlled experiments are in principle impossible. Performing any of the so–called random experiments will not tell us what the "physical probabilities" are, because *there is no such thing as a "physical probability".* The experiment tells us, in a very crude and incomplete way, something about how the initial conditions are varying from one repetition to another.

A much more efficient way of obtaining this information would be to observe the initial conditions directly. However, in many cases this is beyond our present abilities; as in determining the safety and effectiveness of a new medicine. Here the only fully satisfactory approach would be to analyze the detailed sequence of chemical reactions that follow the taking of this medicine, in persons of every conceivable state of health. Having this analysis one could then predict, for each individual patient, exactly what the effect of the medicine will be.

Such an analysis being entirely out of the question at present, the only feasible way of obtaining information about the effectiveness of a medicine is to perform a "random" experiment. No two patients are in exactly the same state of health; and the unknown variations in this factor constitute the variable initial conditions of the experiment, while the sample space comprises the set of distinguishable reactions to the medicine. Our use of probability theory in this case is a standard example of inductive reasoning which amounts to the following:

If the initial conditions of the experiment (*i.e.,* the physiological conditions of the patients who come to us) continue in the future to vary over the same unknown range as they have in the past, then the relative frequency of cures will, in the future, approximate those which we have observed in the past. In the absence of positive evidence giving a reason why there should be some change in the future, *and* indicating in which direction this change should go, we have no grounds for predicting any change in either direction, and so can only suppose that things will continue in more or less the same way. As we observe the relative frequencies of cures and side–effects to remain stable over longer and longer times, we become more and more confident about this conclusion. But this is only inductive reasoning – there is no deductive proof that frequencies in the future will not be entirely different from those in the past.

Suppose now that the eating habits or some other aspect of the life style of the population starts to change. Then the state of health of the incoming patients will vary over a different range than before, and the frequency of cures for the same treatment may start to drift up or down. Conceivably, monitoring this frequency could be a useful indicator that the habits of the population are changing, and this in turn could lead to new policies in medical procedures and public health education.

At this point, we see that the logic invoked here is virtually identical with that of industrial quality control, discussed in Chapter 4. But looking at it in this greater generality makes us see the role of induction in science in a very different way than has been imagined by some philosophers.

Induction Revisited

As we noted in Chapter 9, some philosophers have rejected induction on the grounds that there is no way to prove that it is "right" (theories can never attain a high probability); but this misses the point. The function of induction is to tell us, not which predictions are right, but which predictions are indicated by our present knowledge. If the predictions succeed, then we are pleased and become more confident of our present knowledge; but we have not learned much.

The real role of induction in science was pointed out clearly by Harold Jeffreys (1931, Chapter 1) over sixty years ago; yet to the best of our knowledge no mathematician or philosopher has ever taken the slightest note of what he had to say:

> "A common argument for induction is that induction has always worked in the past and therefore may be expected to hold in the future. It has been objected that this is itself an inductive argument and cannot be used in support of induction. What is hardly ever mentioned is that induction has often failed in the past and that progress in science is very largely the consequence of direct attention to instances where the inductive method has led to incorrect predictions."

Put more strongly, it is only when our inductive inferences are wrong that we learn new things about the real world. For a scientist, therefore, the quickest path to discovery is to examine those situations where it appears most likely that induction from our present knowledge will fail. But those inferences must be our *best* inferences, which make full use of all the knowledge we have. One can always make inductive inferences that are wrong in a useless way, merely by ignoring cogent information.

Indeed, that is just what Popper did. His trying to interpret probability itself as expressing physical causation not only cripples the applications of probability theory in the way we saw in Chapter 3 (it would prevent us from getting about half of all conditional probabilities right because they express logical connections rather than causal physical ones) – it leads one to conjure up imaginary causes while ignoring what was already known about the real physical causes at work. This can reduce our inferences to the level of pre–scientific, uneducated superstition even when we have good data.

Why do physicists see this more readily than others? Because, having created this knowledge of physical law, we have a vested interest in it and want to see it preserved and used. Frequency or propensity interpretations start by throwing away practically all the professional knowledge that we have labored for Centuries to get. Those who have not comprehended this are in no position to discourse to us on the philosophy of science or the proper methods of inference.

But What About Quantum Theory?

Those who cling to a belief in the existence of "physical probabilities" may react to the above arguments by pointing to quantum theory, in which physical probabilities appear to express the most fundamental laws of physics. Therefore let us explain why this is another case of circular reasoning. We need to understand that present quantum theory uses entirely different standards of logic than does the rest of science.

In biology or medicine, if we note that an effect E (for example, muscle contraction, phototropism, digestion of protein) does not occur unless a condition C (nerve impulse, light, pepsin) is present, it seems natural to infer that C is a necessary causative agent for E. Most of what is known in all fields of science has resulted from following up this kind of reasoning. But suppose that condition C does not always lead to effect E; what further inferences should a scientist draw? At this point the reasoning formats of biology and quantum theory diverge sharply.

In the biological sciences one takes it for granted that in addition to C there must be some other causative factor F, not yet identified. One searches for it, tracking down the assumed cause by a process of elimination of possibilities that is sometimes extremely tedious. But persistence pays off; over and over again medically important and intellectually impressive success has been achieved, the conjectured unknown causative factor being finally identified as a definite chemical compound. Most enzymes, vitamins, viruses, and other biologically active substances owe their discovery to this reasoning process.

In quantum theory, one does not reason in this way. Consider, for example, the photoelectric effect (we shine light on a metal surface and find that electrons are ejected from it). The experimental fact is that the electrons do not appear unless light is present. So light must be a causative factor. But light does not always produce ejected electrons; even though the light from a unimode laser is present with absolutely steady amplitude, the electrons appear only at particular times that are not determined by any known parameters of the light. Why then do we not draw the obvious inference, that in addition to the light there must be a second causative factor, still unidentified, and the physicist's job is to search for it?

What is done in quantum theory today is just the opposite; when no cause is apparent one simply postulates that no cause exists – ergo, the laws of physics are indeterministic and can be expressed only in probability form. The central dogma is that the light determines, not whether a photoelectron will appear, but only the probability that it will appear. The mathematical formalism of present quantum theory – incomplete in the same way that our present knowledge is incomplete – does not even provide the vocabulary in which one could ask a question about the real cause of an event.

Biologists have a mechanistic picture of the world because, being trained to believe in causes, they continue to use the full power of their brains to search for them – and so they find them. Quantum physicists have only probability laws because for two generations we have been indoctrinated not to believe in causes – and so we have stopped looking for them. Indeed, any attempt to search for the causes of microphenomena is met with scorn and a charge of professional incompetence and 'obsolete mechanistic materialism'. Therefore, to explain the indeterminacy in current quantum theory we need not suppose there is any indeterminacy in Nature; the mental attitude of quantum physicists is already sufficient to guarantee it.[†]

This point also needs to be stressed, because most people who have not studied quantum theory on the full technical level are incredulous when told that it does not concern itself with causes; and indeed, it does not even recognize the notion of 'physical reality.' The currently taught interpretation of the mathematics is due to Niels Bohr, who directed the Institute for Theoretical Physics in Copenhagen; therefore it has come to be called 'The Copenhagen Interpretation'.

As Bohr stressed repeatedly in his writings and lectures, present quantum theory can answer only questions of the form: "If this experiment is performed, what are the possible results and their probabilities?" It cannot, as a matter of principle, answer any question of the form: "What is really happening when $\cdots$?" Again, the mathematical formalism of present quantum theory, like Orwellian *newspeak*, does not even provide the vocabulary in which one could ask such a question. These points have been explained in some detail in recent articles (Jaynes, 1986d, 1989, 1990a, 1991c).

[†] Here there is a striking similarity to the position of the parapsychologists Soal & Bateman (1954), discussed in Chapter 5. They suggest that to seek a physical explanation of parapsychological phenomena is a regression to the quaint and reprehensible materialism of Thomas Huxley. Our impression is that by 1954 the views of Huxley in biology were in a position of complete triumph over vitalism, supernaturalism, or any other anti–materialistic teachings; for example, the long mysterious immune mechanism was at last understood, and the mechanism of DNA replication had just been discovered. In both cases the phenomena could be described in 'mechanistic' terms so simple and straightforward – templates, geometrical fit, *etc.* – that they would be understood immediately in a machine shop.

We suggest, then, that those who try to justify the concept of 'physical probability' by pointing to quantum theory, are entrapped in circular reasoning, not basically different from that noted above with coins and bridge hands. Probabilities in present quantum theory express the incompleteness of human knowledge just as truly as did those in classical statistical mechanics; only its origin is different.

In classical statistical mechanics, probability distributions represented our ignorance of the true microscopic coordinates – ignorance that was avoidable in principle but unavoidable in practice, but which did not prevent us from predicting reproducible phenomena, just because those phenomena are independent of the microscopic details.

In current quantum theory, probabilities express our own ignorance due to our failure to search for the real causes of physical phenomena – and worse, our failure even to think seriously about the problem. This ignorance may be unavoidable in practice, but in our present state of knowledge we do not know whether it is unavoidable in principle; the "central dogma" simply asserts this, and draws the conclusion that belief in causes, and searching for them, is philosophically naïve. If everybody accepted this and abided by it, no further advances in understanding of physical law would ever be made; indeed, no such advance has been made since the 1927 Solvay Congress in which this mentality became solidified into physics.[‡] But it seems to us that this attitude places a premium on stupidity; to lack the ingenuity to think of a rational physical explanation is to support the supernatural view.

But to many people, these ideas are almost impossible to comprehend because they are so radically different from what we have all been taught from childhood. Therefore let us show how just the same situation could have happened in coin tossing, had classical physicists used the same standards of logic that are now used in quantum theory.

Mechanics Under the Clouds

We are fortunate that the principles of Newtonian mechanics could be developed and verified to great accuracy by studying astronomical phenomena, where friction and turbulence do not complicate what we see. But suppose the Earth were, like Venus, enclosed perpetually in thick clouds. The very existence of an external universe would be unknown for a long time, and to develop the laws of mechanics we would be dependent on the observations we can make locally.

Since tossing of small objects is nearly the first activity of every child, it would be observed very early that they do not always fall with the same side up, and that all one's efforts to control the outcome are in vain. The natural hypothesis would be that it is the volition of the object tossed, not the volition of the tosser, that determines the outcome; indeed, that is the hypothesis that small children make when questioned about this.

Then it would be a major discovery, once coins had been fabricated, that they tend to show both sides about equally often; and the equality appears to get better as the number of tosses increases. The equality of heads and tails would be seen as a fundamental law of physics; symmetric objects have a symmetric volition in falling (as indeed, Cramér and Feller seem to have thought).

With this beginning, we could develop the mathematical theory of object tossing, discovering the binomial distribution, the absence of time correlations, the limit theorems, the combinatorial frequency laws for tossing of several coins at once, the extension to more complicated symmetric objects like dice, etc. All the experimental confirmations of the theory would consist of more and more tossing experiments, measuring the frequencies in more and more elaborate scenarios. From

[‡] Of course, physicists continued discovering new particles and calculation techniques – just as an astronomer can discover a new planet and a new algorithm to calculate its orbit, without any advance in his basic understanding of celestial mechanics.

such experiments, nothing would ever be found that called into question the existence of that volition of the object tossed; they only enable one to confirm that volition and measure it more and more accurately.

Then suppose that someone was so foolish as to suggest that the motion of a tossed object is determined, not by its own volition, but by laws like those of Newtonian mechanics, governed by its initial position and velocity. He would be met with scorn and derision; for in all the existing experiments there is not the slightest evidence for any such influence. The Establishment would proclaim that, since all the observable facts are accounted for by the volition theory, it is philosophically naïve and a sign of professional incompetence to assume or search for anything deeper. In this respect, the elementary physics textbooks would read just like our present quantum theory textbooks.

Indeed, anyone trying to test the mechanical theory would have no success; however carefully he tossed the coin (not knowing what we know) it would persist in showing head and tails about equally often. To find any evidence for a causal instead of a statistical theory, would require control over the initial conditions of launching, orders of magnitude more precise than anyone can achieve by hand tossing. We would continue almost indefinitely, satisfied with laws of physical probability and denying the existence of causes for individual tosses external to the object tossed – just as quantum theory does today – because those probability laws account correctly for everything that we can observe reproducibly with the technology we are using.

But after thousands of years of triumph of the statistical theory, someone finally makes a machine which tosses coins in absolutely still air, with very precise control of the exact initial conditions. Magically, the coin starts giving unequal numbers of heads and tails; the frequency of heads is being controlled partially by the machine. With development of more and more precise machines, one finally reaches a degree of control where the outcome of the toss can be predicted with 100% accuracy. Belief in "physical probabilities" expressing a volition of the coin is recognized finally as an unfounded superstition. The existence of an underlying mechanical theory is proved beyond question; and the long success of the previous statistical theory is seen as due only to the lack of control over the initial conditions of the tossing.

Because of recent spectacular advances in the technology of experimentation, with increasingly detailed control over the initial states of individual atoms [see, for example, Rempe, *et al* (1987); Knight (1987)], we think that the stage is going to be set, before very many more years have passed, for the same thing to happen in quantum theory; a Century from now the true causes of microphenomena will be known to every schoolboy and, to paraphrase Seneca, they will be incredulous that such clear truths could have escaped us throughout the 20'th Century.

More On Coins and Symmetry

Now we go into a more careful, detailed discussion of some of these points, alluding to technical matters that must be explained more fully elsewhere. The rest of this Chapter is not for the casual reader; only the one who wants a deeper understanding than is conveyed by the above simple scenarios. But many of the attacks on Laplace arise from failure to comprehend the following points.

The problems in which intuition compels us most strongly to a uniform probability assignment are not the ones in which we merely apply a principle of "equal distribution of ignorance." Thus, to explain the assignment of equal probabilities to heads and tails on the grounds that we "saw no reason why either face should be more likely than the other," fails utterly to do justice to the reasoning involved. The point is that we have not merely "equal ignorance." We also have *positive knowledge of the symmetry* of the problem; and introspection will show that when this positive knowledge is lacking, so also is our intuitive compulsion toward a uniform distribution. In order to find a respectable mathematical formulation we therefore need to find first a more respectable verbal formulation. We suggest that the following verbalization does do justice to the reasoning, and shows us how to generalize the principle.

"I perceive here two different problems, Having formulated one definite problem – call it P_1 – involving the coin, the operation which interchanges heads and tails transforms the problem into a different one – call it P_2. If I have positive knowledge of the symmetry of the coin, then I know that all relevant dynamical or statistical considerations, however complicated, are exactly the same in the two problems. Whatever state of knowledge I had in P_1, I must therefore have exactly the same state of knowledge in P_2, except for the interchange of heads and tails. Thus, whatever probability I assign to heads in P_1, consistency demands that I assign the *same* probability to tails in P_2.

Now it might be quite reasonable to assign probability 2/3 to heads, 1/3 to tails in P_1; whereupon from symmetry it must be 2/3 to tails, 1/3 to heads in P_2. This might be the case, for example, if P_1 specified that the coin is to be held between the fingers heads up, and dropped just one inch onto a table. Thus symmetry of the coin by no means compels us to assign equal probabilities to heads and tails; the question necessarily involves the other conditions of the problem.

But now suppose the statement of the problem is changed in just one respect; we are no longer told whether the coin is held initially with heads up or tails up. In this case, our intuition suddenly takes over with a compelling force, and tells us that we *must* assign equal probabilities to heads and tails; and in fact, we must do this *regardless of what frequencies have been observed in previous repetitions of the experiment.*

The great power of symmetry arguments lies just in the fact that they are not deterred by any amount of complication in the details. The conservation laws of physics arise in this way; thus conservation of angular momentum for an arbitrarily complicated system of particles is a simple consequence of the fact that the Lagrangian is invariant under space rotations. In current theoretical physics, almost the only known exact results in atomic and nuclear structure are those which we can deduce by symmetry arguments, using the methods of group theory.

These methods could be of the highest importance in probability theory also, if orthodox ideology did not forbid their use. For example, they enable us, in many cases, to extend the principle of indifference to find consistent prior probability assignments in a continuous parameter space Θ, where its use has always been considered ambiguous. The basic point is that a consistent principle for assigning prior probabilities must have the property that it assigns equivalent priors to represent equivalent states of knowledge.

The prior distribution must therefore be invariant under the symmetry group of the problem; and so the prior can be specified arbitrarily only in the so–called "fundamental domain" of the group (Wigner, 1959). This is a subspace $\Theta_0 \subset \Theta$ such that (1) applying two different group elements $g_i \neq g_j$ to Θ_0, the subspaces $\Theta_i \equiv g_i\,\Theta_0$, $\Theta_j \equiv g_j\,\Theta_0$ are disjoint; and (2) carrying out all group operations on Θ_0 just generates the full hypothesis space: $\cup_j\,\Theta_j = \Theta$.

For example, let points in a plane be defined by their polar coordinates (r, α). If the group is the four-element one generated by a 90° rotation of the plane, then any sector 90° wide, such as $(\beta \le \alpha < \beta + \pi/2)$ is a fundamental domain. Specifying the prior in any such sector, symmetry under the group then determines the prior everywhere in the plane.

If the group contains a continuous symmetry operation, the dimensionality of the fundamental domain is less than that of the parameter space; and so the probability density need be specified only on a set of points of measure zero, whereupon it is determined everywhere. If the number of continuous symmetry operations is equal to the dimensionality of the space Θ, the fundamental domain reduces to a single point, and the prior probability distribution is then uniquely determined by symmetry alone, just as it is in the case of an honest coin. Later we shall formalize and generalize these symmetry arguments.

There is still an important constructive point to be made about the power of symmetry arguments in probability theory. To see it, let us go back for a closer look at the coin–tossing problem. The laws of mechanics determine the motion of the coin, as describing a certain trajectory in a twelve–dimensional phase space [three coordinates (q_1, q_2, q_3) of its center of mass, three Eulerian angles (q_4, q_5, q_6) specifying its orientation, and six associated momenta $(p_1, \dots, p_6)$]. The difficulty of predicting the outcome of a toss arises from the fact that very small changes in the location of the initial phase point can change the final results.

Imagine the possible initial phase points to be labelled H or T, according to the final results. Contiguous points labelled H comprise a set which is presumably twisted about in the twelve–dimensional phase space in a very complicated, convoluted way, parallel to and separated by similar T–sets.

Consider now a region R of phase space, which represents the accuracy with which a human hand can control the initial phase point. Because of limited skill, we can be sure only that the initial point is somewhere in R, which has a phase volume

$$\Gamma(R) = \int_R dq_1 \cdots dq_6 \, dp_1 \cdots dp_6$$

If the region R contains both H and T domains, we cannot predict the result of the toss. But what probability should we assign to heads? If we assign equal probability to equal phase volumes in R, this is evidently the fraction $p_H \equiv \Gamma(H)/\Gamma(R)$ of phase volume of R that is occupied by H domains. This phase volume Γ is the "invariant measure" of phase space. The cogency of invariant measures for probability theory will be explained later; for now we note that the measure Γ is invariant under a large group of "canonical" coordinate transformations, and also under the time development, according to the equations of motion. This is Liouville's theorem, fundamental to statistical mechanics; the exposition of Gibbs (1902) devotes the first three Chapters to discussion of it, before introducing probabilities.

Now if we have positive knowledge that the coin is perfectly "honest," then it is clear that the fraction $\Gamma(H)/\Gamma(R)$is very nearly 1/2, and becomes more accurately so as the size of the individual H and T domains become smaller compared to R. Because, for example, if we are launching the coin in a region R where the coin makes fifty complete revolutions while falling, then a one percent change in the initial angular velocity will just interchange heads and tails by the time the coin reaches the floor. Other things being equal, (all dynamical properties of the coin involve heads and tails in the same manner), this should just reverse the final result.

A change in the initial "orbital" velocity of the coin, which results in a one percent change in the time of flight, should also do this (strictly speaking, these conclusions are only approximate, but we expect them to be highly accurate, and to become more so if the changes become less than one percent). Thus, if all other initial phase coordinates remain fixed, and we vary only the initial angular velocity $\dot{\theta}$ and upward velocity $\dot{z}$, the H and T domains will spread into thin ribbons, like the stripes on a zebra. From symmetry, the width of adjacent ribbons must be very nearly equal.

This same "parallel ribbon" shape of the H and T domains presumably holds also in the full phase space.[†] This is quite reminiscent of Gibbs' illustration of fine–grained and coarse–grained probability densities, in terms of the stirring of colored ink in water. On a sufficiently fine scale,

[†] Actually, if the coin is tossed onto a perfectly flat and homogeneous level floor and is not only perfectly symmetrical under the reflection operation that interchanges heads and tails, but also perfectly round, the probability of heads is independent of five of the twelve coordinates, so we have this intricate structure only in a seven–dimensional space. Let the reader for whom this is a startling statement think about it hard, to see why symmetry makes five coordinates irrelevant (they are the two horizontal coordinates of its center of mass, the direction of its horizontal component of momentum, the Eulerian angle for rotation about a vertical axis, and the Eulerian angle for rotation about the axis of the coin).

every phase region is either H or T; the probability of heads is either zero or unity. But on the scale of sizes of the "macroscopic" region R corresponding to ordinary skills, the probability density is the coarse–grained one, which from symmetry must be very nearly 1/2 if we know that the coin is honest.

What if we don't consider all equal phase volumes within R as equally likely? Well, it doesn't really matter if the H and T domains are sufficiently small. "Almost any" probability density which is a smooth, continuous function within R, will give nearly equal weight to the H and T domains, and we will still have very nearly 1/2 for the probability of heads. This is an example of a general phenomenon, discussed by Poincaré, that in cases where small changes in initial conditions produce big changes in the final results, our final probability assignments will be, for all practical purposes, independent of the initial ones.

As soon as we know that the coin has perfect dynamical symmetry between heads and tails – *i.e.*, its Lagrangian function

$$L(q_1 \dots p_6) = (\text{Kinetic energy}) - (\text{Potential energy})$$

is invariant under the symmetry operation that interchanges heads and tails – then we know an exact result. No matter where in phase space the initial region R is located, for every H domain there is a T domain of equal size and identical shape, in which heads and tails are interchanged. Then if R is large enough to include both, we shall persist in assigning probability 1/2 to heads.

But now suppose the coin is biased. The above argument is lost to us, and we expect that the phase volumes of H and T domains within R are no longer equal. In this case, the "frequentist" tells us that there still exists a definite "objective" frequency of heads, $p_H \neq 1/2$ which is a measurable physical property of the coin. Let us understand clearly what this implies. *To assert that the frequency of heads is a physical property only of the coin, is equivalent to asserting that the ratio $v(H)/v(R)$ is independent of the location of region R.* If this were true, it would be an utterly unprecedented new theorem of mechanics, with important implications for physics which extend far beyond coin tossing.

Of course, no such thing is true. From the three specific methods of tossing the coin discussed above which correspond to widely different locations of the region R, it is clear that the frequency of heads will depend very much on how the coin is tossed. Method A uses a region of phase space where the individual H and T domains are large compared to R, so human skill is able to control the result. Method B uses a region where, for a biased coin, the T domain is very much larger than either R or the H domain. Only method C uses a region where the H and T domains are small compared to R, making the result unpredictable from knowledge of R.

It would be interesting to know how to calculate the ratio $v(H)/v(R)$ as a function of the location of R from the laws of mechanics; but it appears to be a very difficult problem. Note, for example, that the coin cannot come to rest until its initial potential and kinetic energy have been either transferred to some other object or dissipated into heat by frictional forces; so all the details of how that happens must be taken into account. Of course, it would be quite feasible to do controlled experiments which measure this ratio in various regions of phase space. But it seems that the only person who would have any use for this information is a professional gambler.

Clearly, our reason for assigning probability 1/2 to heads when the coin is honest is not based merely on observed frequencies. How many of us can cite a single experiment in which the frequency 1/2 was established under conditions we would accept as significant? Yet none of us hesitates a second in choosing the number 1/2. Our real reason is simply common–sense recognition of the *symmetry* of the situation. *Prior information which does not consist of frequencies is of decisive importance in determining probability assignments even in this simplest of all random experiments.*

Those who adhere publicly to a strict frequency interpretation of probability jump to such conclusions privately just as quickly and automatically as anyone else; but in so doing they have violated their basic premise that (probability) $\equiv$ (frequency); and so in trying to justify this choice they must suppress any mention of symmetry, and fall back on remarks about assumed frequencies in random experiments which have, in fact, never been performed.[†]

Here is an example of what one loses by so doing. From the result of tossing a die, we cannot tell whether it is symmetrical or not. But if we know, from direct physical measurements, that the die *is* perfectly symmetrical and we accept the laws of mechanics as correct, then it is no longer plausible inference, but deductive reasoning, that tells us this: *any nonuniformity in the frequencies of different faces is proof of a corresponding nonuniformity in the method of tossing.* The qualitative nature of the conclusions we can draw from the random experiment depend on whether we do or do not know that the die is symmetrical.

This reasoning power of arguments based on symmetry has led to great advances in physics for sixty years; as noted, it is not very exaggerated to say that the only known exact results in mathematical physics are the ones that can be deduced by the methods of group theory from symmetry considerations. Although this power is obvious once noted and it is used intuitively by every worker in probability theory, it has not been widely recognized as a legitimate formal tool in probability theory.[‡]

We have just seen that in the simplest of the random experiments, any attempt to define a probability merely as a frequency involves us in the most obvious logical difficulties as soon as we analyze the mechanism of the experiment. In many situations where we can recognize an element of symmetry our intuition readily takes over and suggests an answer; and of course it is the same answer that our basic desideratum – that equivalent states of knowledge should be represented by equivalent probability assignments – requires for consistency.

But in situations in which we have positive knowledge of symmetry are rather special ones among all those faced by the scientist. How can we carry out consistent inductive reasoning in situations where we do not perceive any clear element of symmetry? This is an open–ended problem because there is no end to the variety of different special circumstances that might arise. As we shall see, the principle of Maximum Entropy gives a useful and versatile tool for many such problems. But in order to give a start toward understanding this, let's go way back to the beginning and consider the tossing of the coin still another time, in a different way.

Independence of Tosses

"When I toss a coin the probability of heads is one half." What do we mean by this statement? Over the past two centuries millions of words have been written about this simple question. A recent exchange (Edwards, 1991) shows that it is still enveloped in total confusion in the minds of some. But by and large, the issue is between the following two interpretations:

> A: "The available information gives me no reason to expect heads rather than tails, or vice versa – I am completely unable to predict which it will be."
>
> B: "If I toss the coin a very large number of times, in the long run heads will occur about half the time – in other words, the frequency of heads will approach 1/2."

We belabor still another time, what we have already stressed many times before: Statement (A) does not describe any property of the coin, but only the robot's *state of knowledge* (or if you prefer,

[†] Or rather, whenever anyone has tried to perform such experiments under sufficiently controlled conditions to be significant, the expected equality of frequencies is *not* observed. The famous experiments of Weldon and Wolf are discussed elsewhere in this work.

[‡] Indeed, L. J. Savage (1962, p. 102) rejects symmetry arguments, thereby putting his system of 'personalistic' probability in the position of recognizing the need for prior probabilities but refusing to admit any formal principles for assigning them.

of ignorance). (B) is, at least by implication, asserting something about the coin. Thus (B) is a very much stronger statement than (A). Note, however, that (A) does not in any way contradict (B); on the contrary, (A) could be a consequence of (B). For if our robot were told that this coin has in the past given heads and tails with equal frequency, this would give it no help at all in predicting the result of the next toss.

Why, then, has interpretation (A) been almost universally rejected by writers on probability and statistics for two generations? There are, we think, two reasons for this. In the first place, there is a widespread belief that if probability theory is to be of any use in applications, we must be able to interpret our calculations in the strong sense of (B). But this is simply untrue, as we have demonstrated throughout the last eight Chapters. We have seen examples of almost all known applications of frequentist probability theory, and many useful problems outside the scope of frequentist probability theory, which are nevertheless solved readily by probability theory as logic.

Secondly, it is another widely held misconception that the mathematical rules of probability theory (the "laws of large numbers") would lead to (B) as a consequence of (A), and this seems to be "getting something for nothing." For, the fact that I know nothing about the coin is clearly not enough to make the coin give heads and tails equally often!

This misconception arises because of a failure to distinguish between the following two statements:

C: "Heads and tails are equally likely on a single toss."

D: "If the coin is tossed N times, each of the 2^N conceivable outcomes is equally likely."

To see the difference between (C) and (D), consider a case where it is known that the coin is biased, but not whether the bias favors heads or tails. Then (C) is applicable but (D) is not. For on this state of knowledge, as was noted already by Laplace, the sequences HH and TT are each somewhat more likely than HT or TH. More generally, our common sense tells us that any unknown influence which favors heads on one toss will likely favor heads on the other toss. Unless our robot has positive knowledge (symmetry of both the coin and the method of tossing) which definitely rules out *all* such possibilities, (D) is not a correct description of his true state of knowledge; it assumes too much.

Statement (D) implies (C), but says a great deal more. (C) says, "I do not know enough about the situation to give me any help in predicting the result of the next throw," while (D) says, "I know that the coin is honest, *and* that it is being tossed in a way which favors neither face over the other, *and* that the method of tossing and the wear of the coin give no tendency for the result of one toss to influence the result of another."

Mathematically, the laws of large numbers require much more than (C) for their derivation. Indeed, if we agree that tossing a coin generates an exchangeable sequence (i.e., the probability that N tosses will yield heads at n specified trials depends only on N and n, not on the order of heads and tails), then application of the de Finetti theorem, as in Chapter 9, shows that the weak law of large numbers holds *only* when (D) can be justified. In this case, it is almost correct to say that the probability assigned to heads is equal to the frequency with which the coin gives heads; because, for any $\epsilon \rightarrow 0$, the probability that the observed frequency $f = (n/N)$ lies in the interval $(1/2 \pm \epsilon)$ tends to unity as $N \rightarrow \infty$. Let us describe this by saying that there exists a *strong connection* between probability and frequency. We analyze this more deeply in Chapter 18.

In most recent treatments of probability theory, the writer is concerned with situations where a strong connection between probability and frequency is taken for granted – indeed this is usually considered essential to the very notion of probability. Nevertheless, the existence of such a strong connection is clearly only an ideal limiting case unlikely to be realized in any real application. For this reason, the laws of large numbers and limit theorems of probability theory can be grossly misleading to a scientist or engineer who naïvely supposes them to be experimental facts, and tries to interpret them literally in his problems. Here are two simple examples:

(1) Suppose there is some random experiment in which you assign a probability p for some particular outcome A. It is important to estimate accurately the fraction f of times A will be true in the next million trials. If you try to use the laws of large numbers, it will tell you various things about f; for example, that it is quite likely to differ from p by less than a tenth of one percent, and enormously unlikely to differ from p by more than one percent. But now, imagine that in the first hundred trials, the observed frequency of A turned out to be entirely different from p. Would this lead you to suspect that something was wrong, and revise your probability assignment for the 101'st trial? If it would, then your state of knowledge is different from that required for the validity of the law of large numbers. You are not sure of the independence of different trials, and/or you are not sure of the correctness of the numerical value of p. Your prediction of f for a million trials is probably no more reliable than for a hundred.

(2) The common sense of a good experimental scientist tells him the same thing without any probability theory. Suppose someone is measuring the velocity of light. After making allowances for the known systematic errors, he could calculate a probability distribution for the various other errors, based on the noise level in his electronics, vibration amplitudes, etc. At this point, a naïve application of the law of large numbers might lead him to think that he can add three significant figures to his measurement merely by repeating it a million times and averaging the results. But, of course, what he would actually do is to repeat some unknown systematic error a million times. It is idle to repeat a physical measurement an enormous number of times in the hope that "good statistics" will average out your errors, because we cannot know the full systematic error. This is the old "Emperor of China" fallacy, discussed elsewhere.

Indeed, unless we know that all sources of systematic error – recognized or unrecognized – contribute less than about one–third the total error, we cannot be sure that the average of a million measurements is any more reliable than the average of ten. Our time is much better spent in designing a new experiment which will give a lower probable error *per trial.* As Poincaré put it, "The physicist is persuaded that one good measurement is worth many bad ones." In other words, the common sense of a scientist tells him that the probabilities he assigns to various errors do not have a strong connection with frequencies, and that methods of inference which presuppose such a connection could be disastrously misleading in his problems.

Then in advanced applications, it will behoove us to consider: How are our final conclusions altered if we depart from the universal custom of orthodox statistics, and relax the assumption of strong connections? Harold Jeffreys showed a very easy way to answer this, as we shall see later. As common sense tells us it must be, the ultimate accuracy of our conclusions is then determined not by anything in the data or in the orthodox picture of things; but rather by our own state of knowledge about the systematic errors. Of course, the orthodoxian will protest that, "We understand this perfectly well; and in our analysis we assume that systematic errors have been located and eliminated." But he does not tell us how to do this, or what to do if – as is the case in virtually every real experiment – they are unknown and so cannot be eliminated. Then all the usual 'asymptotic' rules are qualitatively wrong, and only probability theory as logic can give defensible conclusions.

The Arrogance of the Uninformed

Now we come to a very subtle and important point, which has caused trouble from the start in the use of probability theory. Many of the objections to Laplace's viewpoint which you find in the literature can be traced to the author's failure to recognize it. Suppose we do not know whether a coin is honest, and we fail to notice that this state of ignorance allows the possibility of unknown influences which would tend to favor the same face on all tosses. We say "Well, I don't see any reason why any one of the 2^N outcomes in N tosses should be more likely than any other, so I'll assign uniform probabilities by the principle of indifference."

We would be led to statement (D) and the resulting strong connection between probability and frequency. But this is absurd – in this state of uncertainty, we could not possibly make reliable predictions of the frequency of heads. Statement (D), which is supposed to represent a great deal of positive knowledge about the coin and the method of tossing can also result from *failure* to make proper use of all the available information!

Nothing in our past experience could have prepared us for this; it is a situation without parallel in any other field. In other applications of mathematics, if we fail to use all of the relevant data of a problem, the result will not be that we get an incorrect answer. The result will be that we are unable to get any answer at all. But probability theory cannot have any such built–in safety device, because in principle, the theory must be able to operate no matter what our incomplete information might be.

If we fail to include all of the relevant data or to take into account all the possibilities allowed by the data and prior information, probability theory will still give us a definite answer; and that answer will be the correct conclusion from the information that we actually gave the robot. But that answer may be in violent contradiction to our common–sense judgments which did take everything into account, if only crudely. *The onus is always on the user to make sure that all the information, which his common sense tells him is relevant to the problem, is actually incorporated into the equations and that the full extent of his ignorance is also properly represented.* If you fail to do this, then you should not blame Bayes and Laplace for your nonsensical answers.

We shall see examples of this kind of misuse of probability theory later, in the various objections to the Rule of Succession. It may seem paradoxical that a more careful analysis of a problem may lead to less certainty in prediction of the frequency of heads. However, look at it this way. It is commonplace that in all kinds of questions the fool feels a certainty that is denied to the wise man. The semiliterate on the next bar stool will tell you with absolute, arrogant assurance just how to solve all the world's problems; while the scholar who has spent a lifetime studying their causes is not at all sure how to do this.

In almost any example of inference, a more careful study of the situation, uncovering new facts, can lead us to feel either more certain or less certain about our conclusions, depending on what we have learned. New facts may support our previous conclusions, or they may refute them; we saw some of the subtleties of this in Chapter 5. If our mathematical model failed to reproduce this phenomenon, it could not be an adequate "calculus of inductive reasoning."