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. In addition we have seen that many of the theoretical notions of conventional probability theory, such as sufficient statistics, are readily derivable from the standpoint of probability as logic; and indeed have then a deeper meaning.

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, many expositions *define* probability as meaning 'frequency in a random experiment' rather than as an element of logic. 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 by many to be 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 the conventional way leads to nonsense or disaster. We postpone examination of this pathology to later Chapters, particularly Chapter 17; but the final verdict will be simply: *Just apply the rules of probability theory as logic directly in all problems of inference. This will yield the optimal treatment of the problem, whether or not it could have been solved also by frequentist probability theory.*

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:

† 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.

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.

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 our calculations.* But the very purpose of those calculations 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 specifying m is 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, wheneve 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 many 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 for Popper), 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 we continue to ignore the data (*i.e.*, consider the prior evidence of symmetry to outweigh 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.

† 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 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.

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. While we could make some further progress by elementary methods, any real facility in these calculations will require some more sophisticated mathematical techniques. Since these same techniques will prove equally necessary in most of our later applications, we digress to collect some of the basic mathematical facts needed for them.

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 $\{g_1, g_2 \ldots g_m\}$ 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 gain generated by 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),

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

‡ 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.

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),

$$G = g_j + \sum_{k \neq j} n_k \, g_k \tag{9–14}$$

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

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

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

$$M(n, G) = \sum_{j=1}^{m} M(n-1, G-g_j) \; . \tag{9–19}$$

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

$$\exp(\alpha n + \lambda G) \; . \tag{9–20}$$

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

$$e^{\alpha} = Z(\lambda) \equiv \sum_{j=1}^{m} e^{-\lambda g_j} . \tag{9–21}$$

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:

$$H(n, G) = \int Z^n(\lambda)\, e^{\lambda G}\, h(\lambda)\, d\,\lambda \tag{9–22}$$

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 (9–22) 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:

$$\sum_G e^{-\lambda G}\, M(n, G) = \sum_{\{n_j\}} W(n_1 \ldots n_m)\, \exp(-\lambda \Sigma n_j g_j)\, , \tag{9–23}$$

where the multinomial coefficient

$$W(n_1 \ldots n_m) \equiv \frac{N!}{n_1! \ldots n_m!} \tag{9–24}$$

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

$$\sum_G e^{-\lambda G}\, M(n, G) = Z^n(\lambda)\, . \tag{9–25}$$

Therefore the proper choice of the function $h(\lambda)$ and path of integration in (9–22) is the one that makes (9–22) and (9–25) a Laplace transform pair. To find it, note that the integrand in (9–22) contains a sum of a finite number of terms:

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

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

$$I(u) \equiv \frac{1}{2i} \int_{-iu}^{iu} e^{\lambda(G - G_k)}\, d\lambda = \frac{\sin u(G - G_k)}{G - G_k}\, . \tag{9–27}$$

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

$$\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} Z^n(\lambda)\, e^{\lambda G}\, d\lambda = \sum_k M(n, G_k)\, \delta(G - G_k) \tag{9–28}$$

and of course (9–25) can be written more explicitly as

$$Z^n(\lambda) = \int e^{-\lambda G}\, q(G)\, dG \tag{9–29}$$

where

$$q(G) \equiv \sum_k M(n, G_k)\, \delta(G - G_k)\ . \tag{9–30}$$

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

We consider the use of this presently, but note first that in many cases (9–25) is all we need to solve combinatorial problems.

Equation (9–25) 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

$$g_j \equiv \ \delta(i, 1) \tag{9–31}$$

then the total G is just the first sample number:

$$G = \sum n_j g_j = n_1\ . \tag{9–32}$$

The partition function (9–21) is then

$$Z(\lambda) = e^{-\lambda} + m - 1 \tag{9–33}$$

and from Newton's binomial expansion,

$$Z^n(\lambda) = \sum_{s=0}^{n} \binom{n}{s}\, e^{-\lambda s}\, (m-1)^{n-s}\ . \tag{9–34}$$

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

$$M(n, G) = M(n, n_1) = \binom{n}{n_1}\, (m-1)^{n-n_1}\ . \tag{9–35}$$

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) $\times$ (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

† 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. Further comments on this are in Appendix B.

complicated problems; and even in this example the partition function method, once understood, is easier to use.

In the choice (9–31) 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 (9–31) by

$$g_j = \left\{ \begin{array}{ll} 1, & j \in S \\ 0, & \text{otherwise} \end{array} \right\} \tag{9–36}$$

and Equations (9–32)–(9–35) are generalized as follows. G is now the total number of successes, called traditionally r:

$$G = \sum_{j=1}^{m} n_j \, g_j = r \tag{9–37}$$

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

$$Z(\lambda) = s\, e^{-\lambda} + m - s \tag{9–38}$$

from which

$$Z^n(\lambda) = \sum_{r=0}^{n} \binom{n}{r} s^r e^{-\lambda r} (m - s)^{n-r} \tag{9–39}$$

and

$$M(n, G) = M(n, r) = \binom{n}{r} \; s^r \; (m - s)^{n-r} \; . \tag{9–40}$$

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

$$P(G = r|I_0) = \binom{n}{r} \, p^r \, (1 - p)^{n-r} \; , \qquad 0 \le r \le n \tag{9–41}$$

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 the binomial distribution 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 all the usual probability distributions (Multinomial, Poisson, Gamma, Gaussian, Chi–squared *etc.*) that conventional probability theory has been obtaining from the assumption of strict independence of different trials. All of the mathematical results that Feller and other sampling theorists find from the model of independent repetitions of a random experiment, 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*.

Then, since the poorly informed robot is unable to do inductive reasoning, we begin to understand why conventional probability theory is unable to account for 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, as we shall see in Chapters 16, 17, 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 (9–39) becomes

$$\frac{Z^n(\lambda)}{Z^n(0)} = \sum_{r=0}^{n} b(r|n,p)\, z^r \tag{9–42}$$

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 (9–22) 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 formnulas. We shall see many 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 (9–22) 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 (9–22) 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 (9–22) 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 (9–23) and (9–24) we see that the multiplicity of ways in which the total G can be realized can be written as

$$M(n,G) = \sum_{\{n_j\}} W(n_1 \cdots n_m) \tag{9–43}$$

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

$$\sum n_j = n \;, \qquad \sum n_j g_j = G \;. \tag{9–44}$$

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

$$\sum_{j=1}^{m} k_j g_j = 0 \tag{9–45}$$

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 (9–45) 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 (9–43):

$$M(n, G) = W(n_1 \cdots n_m) \tag{9–46}$$

and

$$M(n-1, G-g_j) = W(n'_1 \cdots n'_m) \tag{9–47}$$

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

$$p(r_k = j|G, n, I_0) = \frac{W(n'_1 \cdots n'_m)}{W(n_1 \cdots n_m)} = \frac{(n-1)!}{n!} \frac{n_j!}{(n_j-1)!} = \frac{n_j}{n} \tag{9–48}$$

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:

$$P(r_k = j|n_j, I_0) = \frac{n_j}{(total\ number\ of\ digits)} \tag{9–49}$$

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 our robot, even if poorly informed, is nevertheless producing some standard results that current conventional treatments all assure us are correct. Indeed, virtually all of current probability theory – in which this strong connection between probability and frequency is taken for granted for all probabilities – can be derived from the poorly informed robot's results.

This only makes the mystery seem deeper and raises new questions: do those standard results also "fail to take the hint?" Or, perhaps those results follow from more than one state of knowledge. So let us note another viewpoint.

We are taught what is called the "law" of radioactive decay – a peculiar term, since nobody has ever seen an accurate verification of the law. The best one can say is that we seldom see large departures from it, except in special circumstances where this was to be expected.

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.

We have noted, in (5–4) and (5–5), a different asymptotic formula in terms of entropy. In Appendix E it is shown that the entropy formula is far more accurate.

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 COMNG 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.‡

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

† 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 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.

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 $n(25) - n(26)$ 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 ($8 \le x < 9$); *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; ($7 < x \le 8$).

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 ($24 \le x < 25$). 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. Had we not done the induction, we would not know what to look for in order to learn this. So induction is most valuable to a scientist just when it turns out to be wrong. Those who criticize induction on the grounds that it might not be right, could not possibly be more mistaken. They are victims of the Mind Projection Fallacy in supposing that probability distributions must describe reality in order to be useful; from the way we construct them when we perform induction, it is clear that they describe only our 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 now 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 and quite similar 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 following just 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; but it indicates it with a high probability *within the hypothesis space being used*, whereupon direct tests can confirm it unequivocally. 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, and is considered in more detail in Chapter 16, 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.†

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. 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.‡

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. 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.⋆

† 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.

‡ 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.

These documentaries always have very beautiful photography, and they deserve commentaries that make sense.