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.

Most of probability theory, however, 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 relationship between these conventional treatments and the theory being developed here.

Some of these relationships have been seen already; in the preceding 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 probability theory as logic, 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 better 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.

9.1 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 currently 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 the 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 factors 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 factors which vary in an uncontrolled way are often called *random*, a term which we shall usually avoid, because in current English usage it carries some very wrong connotations.[1] We should call them, rather, *irreproducible by the experimental technique used*. They might become reproducible by an improved technique; indeed, the progress of all areas of experimental science involves the continual development of more powerful techniques that exert finer control over conditions, making more effects reproducible. Once a phenomenon becomes reproducible, as has happened in molecular biology, it emerges from the cloud of speculation and fantasy to become a respectable part of ‘hard’ science.

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

[1] 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.

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 an 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 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 persons with reasonably good 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 results per trial *that we shall take into account in the present calculation*. 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. In any event, we are concerned with two different sample spaces; the space S for a single trial, consisting of m points, and the extension space

$$S^n = S \otimes S \otimes \cdots \otimes S, \tag{9.1}$$

the direct product of n copies of S, which is the sample space for the experiment as a whole. For clarity, we use the word ‘result’ for a single trial referring to space S, while ‘outcome’ refers to the experiment as a whole, defined on space S^n. Thus, one outcome consists of the enumeration of n results (including their order if the experiment is conducted in such a way that an order is defined). Then we may say that the number of results *being considered in the present calculation* is m, while the number of outcomes being considered is $N = m^n$.

Denote the result of the ith trial by r_i $(1 \le r_i \le m,\ 1 \le i \le n)$. 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 function of the r_i:

$$P(D|I) = p(r_1 \ldots r_n) \tag{9.2}$$

satisfying the sums over all possible data sets

$$\sum_{r_1=1}^{m} \sum_{r_2=1}^{m} \cdots \sum_{r_n=1}^{m} p(r_1 \ldots r_n) = 1. \tag{9.3}$$

As a convenience, since the r_i are non-negative integers, we may regard them as digits (*modulo* m) in a number R expressed in the base m number system; $0 \le R \le N - 1$. Our robot, however poorly informed it may be about the real world, is an accomplished manipulator of numbers, so we may instruct it to communicate with us in the base m number system instead of the decimal (base ten) system that you and I were 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, which can be indexed (1 to 1296). Then to indicate the outcome that is designated as number 836 in the decimal system, the robot notes that

$$836 = (3 \times 6^3) + (5 \times 6^2) + (1 \times 6^1) + (2 \times 6^0) \tag{9.4}$$

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

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 six system the individual digits r_i have meaning only *modulo* 6, the display $5024 \equiv 5624$ 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 with the robot in the base m number system, whereupon each number displayed will have exactly n digits, and for us the ith digit will represent, *modulo* m, the result of the ith 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. (2.86) that, depending on details of the information I, many different probability assignments (9.2) might be appropriate; consider first the obvious simplest case of all.

9.2 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 rational inferences about the experiment; and so we have. Nevertheless, it is important that we understand the surprising consequences of neglecting that information.

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 some 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 (2.95) 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 $S' \subset S^n$ and false on the complementary subset $S^n - S'$ will, by the rule (2.99), then be assigned the probability

$$P(A|I_0) = \frac{M(n, A)}{N}, \tag{9.5}$$

where $M(n, A)$ is the multiplicity of A (number of points of S^n on which A is true). This trivial looking result summarizes everything the robot can say on the prior information I_0, and it illustrates again that, whenever they are relevant to the problem, connections between probability and frequency appear automatically, as mathematical consequences of our rules.

Consider n tosses of a die, $m = 6$; the probability (9.2) of any completely specified outcome is

$$p(r_1 \dots r_n|I_0) = \frac{1}{6^n}, \qquad 1 \le r_i \le 6, \quad 1 \le i \le n. \tag{9.6}$$

Then 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

$$A(r_2, \dots, r_n) \equiv r_1 = 3 \text{ and the remaining digits are } r_2, \dots, r_n \tag{9.7}$$

are mutually exclusive, and so (2.85) applies:

$$\begin{aligned} P(r_1 = 3|I_0) &= \sum_{r_2=1}^{6} \cdots \sum_{r_n=1}^{6} p(3\, r_2 \dots r_n|I_0) \\ &= 6^{n-1} p(r_1 \dots r_n|I_0) \\ &= 1/6. \end{aligned} \tag{9.8}$$

(Note that '$r_1 = 3$' is a proposition, so by our notational rules in Appendix B we are allowed to put it in a formal probability symbol with capital P.) But by symmetry, if we had asked for the probability that any specified (ith) toss gives any specified (kth) result, the calculation would have been the same:

$$P(r_i = k|I_0) = 1/6, \qquad 1 \le k \le 6, \quad 1 \le i \le n. \tag{9.9}$$

Now, what is the probability that the first toss gives k 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

$$\begin{aligned} P(r_1 = k, r_2 = j|I_0) &= \sum_{r_3=1}^{6} \cdots \sum_{r_n=1}^{6} p(k, j, r_3 \dots r_n|I_0) \\ &= 6^{n-2} p(r_1 \dots r_n|I_0) = 1/6^2 \\ &= 1/36, \end{aligned} \tag{9.10}$$

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

$$p(r_i r_j r_k|I_0) = 1/6^3 = 1/216, \tag{9.11}$$

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 three 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 (2.63):

$$p(r_2|r_1I_0) = \frac{p(r_1r_2|I_0)}{p(r_1|I_0)}, \tag{9.12}$$

or, using (9.9) and (9.10),

$$p(r_2|r_1I_0) = \frac{1/36}{1/6} = 1/6 = p(r_2|I_0). \tag{9.13}$$

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 (9.11) the same result:

$$p(r_3|r_1r_2I_0) = \frac{p(r_3r_1r_2|I_0)}{p(r_1r_2|I_0)} = \frac{1/216}{1/36} = 1/6 = p(r_3|I_0). \tag{9.14}$$

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.

9.3 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, as in Chapters 1 and 2, 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, and the nemesis of some philosophers beginning with David Hume (1739, 1777) in the 18th century. 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 providing a logical connection between 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 $67 \pm 3\%$ 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 scenarios. 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% 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.

9.4 Are there general inductive rules?

What is shown by (9.13) and (9.14) 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 – physical or logical – 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).

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.

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.

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 labored for years to discover. 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 (3.14), where we examined the hypergeometric distribution. But on second glance it was doing 'reverse induction'; the more red balls that 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 (9.12) but with the hypergeometric distribution $h(r|N, M, n)$ of (3.22):

$$P(H_4|DI_4) = \frac{h(4|100, 50, 4)}{h(3|100, 50, 3)} = \frac{0.05873}{0.12121} = 0.4845 < \frac{1}{2}. \tag{9.15}$$

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 conclusions 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 some system of parameters.

Nevertheless, such a classification was attempted by the philosopher R. Carnap (1891–1970), who found (Carnap, 1952) a continuum of rules identified 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 18th century (the 'rule of succession' and its generalizations) that had been rejected as metaphysical nonsense by statisticians and philosophers.[2]

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), de Finetti (1937) and Harold Jeffreys (1939), of which Carnap seemed unaware.

Carnap was 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 he suffered from one of the standard occupational diseases of philosophers; his exposition wanders off into abstract symbolic logic without ever considering a specific real example. So he 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; there is no 'general inductive rule'. Yet neither the term 'prior information' nor the concept ever appears in Carnap's exposition.

[2] Carnap (1952, p. 35), like Venn (1866), 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 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.

9.5 Multiplicity factors

In spite of the formal simplicity of (9.5), 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

$$6^{12} = 2.18 \times 10^9, \tag{9.16}$$

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 of the order of 10^{10} years, or 3×10^{17} seconds. Thus, in 30 tosses of a die, the number of possible outcomes ($6^{30} = 2.21 \times 10^{23}$) is about equal to the number of microseconds in the age of the universe. Yet we shall be particularly interested in evaluating quantities like (9.5) 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. While we could make some further progress by elementary methods, any real facility in these calculations requires some more sophisticated mathematical techniques. We digress to collect some of the basic mathematical facts needed for them. These were given, for the most part, by Laplace, J. Willard Gibbs, and Claude Shannon. In view of the large numbers, there turn out to be extremely good approximations which are easy to calculate.

A large class of problems may be fit into the following scheme, for which we can indicate the exact calculation that should, in principle, be done. Let $\{g_1, g_2, \ldots, g_m\}$ be any set of m finite real numbers. For concreteness, one may think of g_j as the 'value' or the 'gain' of observing the jth 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 such as $g_1 + g_2$ are to be, like sums of pennies, meaningful to us. We could, equally well, make it more abstract by saying simply that we are concerned with predicting linear functions of the n_j. The total amount of G generated by the experiment is then

$$G = \sum_{i=1}^{n} g(r_i) = \sum_{j=1}^{m} n_j g_j, \tag{9.17}$$

where the sample number n_j is the number of times the jth result occurred. If we ask the robot for the probability for obtaining this amount, it will answer, from (9.5),

$$p(G|n, I_0) = f(G|n, I_0) = \frac{M(n, G)}{N}, \tag{9.18}$$

where $N = m^n$ and $M(n, G)$ 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 needed to define an outcome – because we want to allow this to vary). Many probabilities are determined by this multiplicity factor, in its dependence on n and G.

9.6 Partition function algorithms

Expanding $M(n, G)$ according to the result of the nth trial gives the recursion relation

$$M(n, G) = \sum_{j=1}^{m} M(n-1, G-g_j). \tag{9.19}$$

For small n, a computer could apply this n times for direct evaluation of $M(n, G)$, but this would be impractical for very large n. Equation (9.19) is a linear difference equation with constant coefficients in both n and G, so it must have elementary solutions of exponential form:

$$\exp\{\alpha n + \lambda G\}. \tag{9.20}$$

On substitution into (9.19), we find that this is a solution of the difference equation if α and λ are related by

$$\exp\{\alpha\} = Z(\lambda) \equiv \sum_{j=1}^{m} \exp\{-\lambda g_j\}. \tag{9.21}$$

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:

$$H(n, G) = \int \mathrm{d}\lambda\, Z^n(\lambda) \exp\{\lambda G\} h(\lambda) \tag{9.22}$$

is, from linearity, 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 $G = \sum n_j g_j$, the values that are possible results of n trials. Further elaboration of (9.22) leads to analytical methods of calculation that will be used in the advanced applications in the later chapters; but for the present let us note the remarkable things that can be done just with algebraic methods.

Equation (9.22) has the form of an inverse Laplace transform. To find the discrete Laplace transform of $M(n, G)$ 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:

$$\sum_G \exp\{-\lambda G\}\, M(n, G) = \sum_{n_j \in U} W(n_1, \ldots, n_m) \exp\left\{-\lambda \sum n_j g_j\right\}, \tag{9.23}$$

where the multinomial coefficient

$$W(n_1, \ldots, n_m) \equiv \frac{n!}{n_1! \cdots n_m!} \tag{9.24}$$

is the number of outcomes which lead to the sample numbers $\{n_j\}$. If $x_j^{n_j} = \exp\{-n_j g_j\}$ then $\exp\{-\sum_j^m n_j g_j\} = x_1^{n_1} x_2^{n_2} \ldots\ x_m^{n_m}$. The multinomial expansion is defined by

$$(x_1 + \cdots + x_m)^n = \sum_{n_j \in U} W(n_1, \ldots, n_m) x_1^{n_1} \ldots\ x_m^{n_m}. \tag{9.25}$$

In (9.23) we sum over the 'universal set' U, defined by

$$\left\{U : n_j \geq 0, \quad \sum_{j=1}^m n_j = n\right\}, \tag{9.26}$$

which consists of all possible sample numbers in n trials. But, comparing (9.23) with (9.25), this is just

$$\sum_G \exp\{-\lambda G\} M(n, G) = Z^n(\lambda). \tag{9.27}$$

Equation (9.27) 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 nth 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.

9.6.1 Solution by inspection

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

$$g_j \equiv \delta(j, 1), \tag{9.28}$$

then the total G is just the first sample number:

$$G = \sum_j n_j g_j = n_1. \tag{9.29}$$

The partition function (9.21) is then

$$Z(\lambda) = \exp\{-\lambda\} + m - 1 \tag{9.30}$$

and, from Newton's binomial expansion,

$$Z^n(\lambda) = \sum_{s=0}^{n} \binom{n}{s} \exp\{-\lambda s\}(m-1)^{n-s}. \tag{9.31}$$

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

$$M(n, G) = M(n, n_1) = \binom{n}{n_1}(m-1)^{n-n_1}. \tag{9.32}$$

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

In the choice (9.28) we separated off the trial result $j = 1$ for special attention. More generally, suppose we separate the m trial results comprising the sample space S 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 (9.28) by

$$g_j = \begin{cases} 1 & j \in S' \\ 0 & \text{otherwise,} \end{cases} \tag{9.33}$$

and (9.29)–(9.32) are generalized as follows. G is now the total number of successes, called traditionally r:

$$G = \sum_{j=1}^{m} n_j g_j \equiv r, \tag{9.34}$$

and the partition function now becomes

$$Z(\lambda) = s\exp\{-\lambda\} + m - s, \tag{9.35}$$

from which

$$Z^n(\lambda) = \sum_{r=0}^{n} \binom{n}{r} s^r \exp\{-\lambda r\}(m-s)^{n-r}, \tag{9.36}$$

and so the coefficient of $\exp\{-\lambda r\}$ is

$$M(n, G) = M(n, r) = \binom{n}{r} s^r (m-s)^{n-r}. \tag{9.37}$$

From (9.18), the poorly informed robot's probability for r successes is therefore

$$P(G = r|I_0) = \binom{n}{r} p^r (1-p)^{n-r}, \qquad 0 \le r \le n, \tag{9.38}$$

where $p = s/m$. But this is just the binomial distribution $b(r|n, p)$, whose derivation cost us so much conceptual agonizing in Chapter 3. There we found the binomial distribution (3.86) as the limiting form in drawing from an infinitely large urn, and again as a randomized approximate form (3.92) 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! 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 for any result is to be the same at each trial, *strictly independent 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.

Exercise 9.2. Derive the multinomial distribution found in Chapter 3, Eq. (3.77), as a generalization or extension of our derivation of (9.38).

Then all the usual probability distributions of sampling theory (Poisson, gamma, Gaussian, chi-squared, etc.) will follow as limiting forms of these. 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 problems. In other words, *frequentist probability theory is, functionally, just 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 has trouble with it. Until we learn how to introduce some kind of logical connection 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'.

Frequentist probability theory seems to be stuck with independent trials because it lays great stress on limit theorems, and examination of them shows that their derivation 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, not only limit theorems, but virtually all of the sampling distributions for estimators, on which orthodox statistics depends, would be incorrect.

Here the poorly informed robot would seem to have the tactical advantage; for all those limit theorems and sampling distributions for estimators are valid exactly on information I_0. There is another important difference; in conventional probability theory that 'independence' is held to mean causal physical independence; but how is one to judge this as a property of the real world? We have seen no discussion of this in the orthodox literature. To the robot it means logical independence, a stronger condition, but one that makes its calculations cleaner and simpler.

Solution by inspection of $Z^n(\lambda)$ has the merit that it yields exact results. However, only relatively simple problems can be solved in this way. We now note a much more powerful algebraic method.

9.7 Entropy algorithms

We return to the problem of calculating multiplicities as in (9.18)–(9.37), but in a little more general formulation. Consider a proposition $A(n_1, \dots, n_m)$ which is a function of the sample numbers n_j; it is defined to be true when $(n_1, \dots, n_m)$ are in some subset $R \in U$, where U is the universal set (9.26), and false when they are in the complementary set $\overline{R} = U - R$. If A is linear in the n_j, then it is the same as our G in (9.17). The multiplicity of A (number of outcomes for which it is true) is

$$M(n, A) = \sum_{n_j \in R} W(n_1, \dots, n_m), \tag{9.39}$$

where the multinomial coefficient W was defined in (9.24).

How many terms $T(n, m)$ are in the sum (9.39)? This is a well-known combinatorial problem for which the reader will easily find the solution[3]

$$T(n, m) = \binom{n + m - 1}{n} = \frac{(n + m - 1)!}{n!(m - 1)!}, \tag{9.40}$$

and we note that, as $n \to \infty$,

$$T(n, m) \sim \frac{n^{m-1}}{(m - 1)!}. \tag{9.41}$$

The number of terms grows as a finite [$(m - 1)$th] power of n (as can be seen intuitively by thinking of the n_j as Cartesian coordinates in an m-dimensional space and noting the geometrical meaning of the conditions (9.26) defining U). Denote the greatest term in the

[3] Physicists will recognize $T(n, m)$ as the 'Bose–Einstein multiplicity factor' of statistical mechanics (the number of linearly independent quantum states which can be generated by putting n Bose–Einstein particles into m single-particle states). Finding $T(n, m)$ is the same combinatorial problem.

region R by

$$W_{\max} \equiv \text{Max}_R W(n_1, \ldots, n_m). \tag{9.42}$$

Then the sum (9.39) cannot be less than $W_{\max}$, and the number of terms in (9.39) cannot be greater than $T(n, m)$, so

$$W_{\max} \leq M(n, A) \leq W_{\max} T(n, m) \tag{9.43}$$

or

$$\frac{1}{n} \log(W_{\max}) \leq \frac{1}{n} \log M(n, A) \leq \frac{1}{n} \log(W_{\max}) + \frac{1}{n} \log T(n, m). \tag{9.44}$$

But as $n \to \infty$, from (9.41), we have

$$\frac{1}{n} \log T(n, m) \to 0, \tag{9.45}$$

and so

$$\frac{1}{n} \log M(n, A) \to \frac{1}{n} \log(W_{\max}). \tag{9.46}$$

The multinomial coefficient W grows so rapidly with n that in the limit the single maximum term in the sum (9.39) dominates it. The logarithm of T grows less rapidly than n, so in the limit it makes no difference in (9.44).

Then how does $\log(W/n)$ behave in the limit? The limit we want is the one in which the sample frequencies $f_j = n_j/n$ tend to constants; in other words, the limit as $n \to \infty$ of

$$\frac{1}{n} \log \left[\frac{n!}{(nf_1)! \cdots (nf_m)!} \right] \tag{9.47}$$

as the f_j are held constant. But, from the Stirling asymptotic approximation,

$$\log(n!) \sim n \log(n) - n + \log \sqrt{2\pi n} + O\left(\frac{1}{n}\right), \tag{9.48}$$

we find that, in the limit, $\log(W/n)$ tends to a finite constant value independent of n:

$$\frac{1}{n} \log(W) \to H \equiv -\sum_{j=1}^{m} f_j \log(f_j), \tag{9.49}$$

which is just what we call the *entropy* of the frequency distribution $\{f_1, \ldots, f_m\}$. We have the result that, for very large n, if the sample frequencies f_j tend to constants, the multiplicity of A goes into a surprisingly simple expression:

$$M(n, A) \sim \exp\{nH\}, \tag{9.50}$$

in the sense that the ratio of the two sides in (9.50) tends to unity (although their difference does not tend to zero; but they are both growing so rapidly that this makes no percentage difference in the limit). From (9.46) it is understood that in (9.50) the frequencies $f_j = n_j/n$ to be used in H are the ones which maximize H over the region R for which A is defined.

We now see what was not evident before; that this multiplicity is to be found by determining the *frequency* distribution $\{f_1, \ldots, f_m\}$ which has maximum entropy subject to whatever constraints define R.[4]

It requires some thought and analysis to appreciate what we have in (9.50). Note first that we now have the means to complete the calculations which require explicit values for the multiplicities $M(n, G)$. Before proceeding to calculate the entropy, let us note briefly how this will go. If A is linear in the n_j, then the multiplicity (9.50) is also equal asymptotically to

$$M(n, G) = \exp\{nH\}, \tag{9.51}$$

and so the probability for realizing the total G is, from (9.18) ,

$$p(G|n, I_0) = m^{-n} \exp\{nH\} = \exp\{-n(H_0 - H)\}, \tag{9.52}$$

where $H_0 = \log(m)$ is the absolute maximum of the entropy, derived below in (9.74). Often, the quantity most directly relevant is not the entropy, but the difference between the entropy and its maximum possible value; this is a direct measure of how strong are the constraints R. For many purposes it would have been better if entropy had been defined as that difference; but the historical precedence would be very hard to change now. In any event, (9.52) has some deep intuitive meaning that we develop in later chapters.

Let us note the effect of acquiring new information; now we learn that a specified trial yielded the amount g_j. This new information changes the multiplicity of A because now the remaining $(n - 1)$ trials must have yielded the total amount $(G - g_j)$, and the number of ways this could happen is $M(n - 1, G - g_j)$. Also, the frequencies are slightly changed, because one trial that yielded g_j is now absent from the counting. Instead of $f_k = n_k/n$ in (9.18), we now have the frequencies $\{f'_1, \ldots, f'_m\}$, where

$$f'_k = \frac{n_k - \delta_{jk}}{n - 1}, \qquad 1 \le k \le m, \tag{9.53}$$

or writing $f'_k = f_k + \delta f_k$, the change is

$$\delta f_k = \frac{f_k - \delta_{jk}}{n - 1}, \tag{9.54}$$

which is exact; as a check, note that $\sum f'_k = 1$ and $\sum \delta f_k = 0$, as it should be.

This small change in frequencies induces a small change in the entropy; writing the new value as $H' = H + \delta H$, we find

$$\delta H = \sum_k \frac{\partial H}{\partial f_k} \delta f_k + O\left(\frac{1}{n^2}\right) = \left[\frac{H + \log(f_j)}{n - 1}\right] + O\left(\frac{1}{n^2}\right), \tag{9.55}$$

[4] We now also see that, not only is the notion of entropy inherent in probability theory independently of the work of Shannon, the maximum entropy principle is also, at least in this case, derivable directly from the rules of probability theory without additional assumptions.

and so

$$H' = \frac{nH + \log(f_j)}{n-1} + O\left(\frac{1}{n^2}\right). \tag{9.56}$$

Then the new multiplicity is, asymptotically,

$$M(n-1, G-g_j) = \exp\{(n-1)H'\} = f_j \exp\{nH\}\left[1 + O\left(\frac{1}{n}\right)\right]. \tag{9.57}$$

The asymptotic forms of multiplicity are astonishingly simple compared with the exact expressions. This means that, contrary to first appearances when we noted the enormous size of the extension set S^n, the large n limit is by far the *easiest* thing to calculate when we have the right mathematical machinery. Indeed, the set S^n has disappeared from our considerations; the remaining problem is to calculate the f_k that maximize the entropy (9.49) over the domain R. But this is a problem that is solved on the sample space S of a single trial!

The probability for getting the total gain G is changed from (9.52) to

$$p(G|r_i = j, nI_0) = \frac{M(n-1, G-g_j)}{m^{n-1}}, \tag{9.58}$$

and, given only I_0, the prior probability for the event $r_i = j$ is, from (9.5),

$$p(r_i = j|nI_0) = \frac{1}{m}. \tag{9.59}$$

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

$$p(r_i = j|GnI_0) = p(r_i = j|nI_0)\frac{p(G|r_i = j, nI_0)}{p(G|nI_0)}, \tag{9.60}$$

or

$$p(r_i = j|GnI_0) = \frac{1}{m}\frac{[M(n-1, G-g_j)/m^{n-1}]}{[M(n, G)/m^n]} = \frac{M(n-1, G-g_j)}{M(n, G)} = f_j. \tag{9.61}$$

Knowledge of G therefore changes the robot's probability for the jth result from the uniform prior probability $1/m$ to the observed frequency f_j of that result. Intuition might have expected this connection between probability and frequency to appear eventually, but it may seem surprising that this requires only that the total G be known. Note, however, that specifying G determines the maximum entropy frequency distribution $\{f_1, \ldots, f_m\}$, so there is no paradox here.

Exercise 9.3. Extend this result to derive the joint probability

$$p(r_i = j, r_s = t|GnI_0) = M(n-2, G-g_j-g_t)/M(n, G) \tag{9.62}$$

as a ratio of multiplicities and give the resulting probability. Are the trials still independent, or does knowledge of G induce correlations between different trials?

These results show the gratifyingly simple and reasonable things that the poorly informed robot can do. In conventional frequentist probability theory, these connections are only postulated arbitrarily; the poorly informed robot derives them as consequences of the rules of probability theory.

Now we return to the problem of carrying out the entropy maximization to obtain explicit expressions for the entropies H and frequencies f_j.

9.8 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 (9.28) and (9.29) we see that the multiplicity of ways in which the total G can be realized can be written as

$$M(n, G) = \sum_{\{n_j\}} W(n_1, \ldots, n_m), \tag{9.63}$$

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

$$\sum n_j = n, \qquad \sum n_j g_j = G. \tag{9.64}$$

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

$$\sum_{j=1}^{m} k_j g_j = 0, \tag{9.65}$$

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 close by choosing r, s large enough). Likewise, we shall call the numbers $(g_1, \ldots, g_m)$ *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 (9.65) implies that all $k_j = 0$:

$$n_j = n'_j, \qquad 1 \le j \le m, \tag{9.66}$$

so if the $\{g_1, \ldots, g_m\}$ 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 (9.63):

$$M(n, G) = W(n_1, \ldots, n_m) \tag{9.67}$$

and

$$M(n-1, G-g_j) = W(n'_1, \ldots, n'_m), \tag{9.68}$$

where, necessarily, $n'_i = n_i - \delta_{ij}$. Then the exact result (9.61) reduces to

$$p(r_k = j|GnI_0) = \frac{W(n'_1, \ldots, n'_m)}{W(n_1, \ldots, n_m)} = \frac{(n-1)!}{n!} \frac{n_j!}{(n_j - 1)!} = \frac{n_j}{n}. \tag{9.69}$$

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 $\{r_1, \ldots, r_n\}$ equal to j) but does not know at which trials the jth result occurred (i.e., which digits are equal to j), it can apply James Bernoulli's rule (9.18) directly:

$$P(r_k = j|n_j I_0) = \frac{n_j}{\text{(total number of digits)}}. \tag{9.70}$$

Again, the *probability* for 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 $1/\sqrt{n}$, 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 18th 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.

9.9 Entropy maximization

The above derivation (9.50) of $M(n, A)$ is valid for a proposition A that is defined by some arbitrary function of the sample numbers n_j. In general, one might need many different algorithms for this maximization. But in the case $A = G$, where we are concerned with a linear function $G = \sum n_j g_j$, the domain R is defined by specifying just the *average* of G over the n trials:

$$\overline{G} = \frac{G}{n} = \sum_{j=1}^{m} f_j g_j, \tag{9.71}$$

which is also an average over the frequency distribution. Then the maximization problem has a solution that was given once and for all by J. Willard Gibbs (1902) in his work on statistical mechanics.

It required another lifetime for Gibbs' algorithm to be generally appreciated; for 75 years it was rejected and attacked by some because, for those who thought of probability as a real physical phenomenon, it appeared arbitrary. Only through the work of Claude Shannon (1948) was it possible to understand what the Gibbs' algorithm was accomplishing. This was pointed out first in Jaynes (1957a) in suggesting a new interpretation of statistical mechanics (as an example of logical inference rather than as a physical theory), and this led rather quickly to a generalization of Gibbs' equilibrium theory to nonequilibrium statistical mechanics.

In Chapter 11 we set down the complete mathematical apparatus generated by the maximum entropy principle; for the present it will be sufficient to give the solution for the case at hand. An inequality given by Gibbs leads to an elegant solution to our maximization problem.

Let $\{f_1, \ldots, f_m\}$ be any possible frequency distribution on m points, satisfying ($f_j \geq 0$, $\sum_j f_j = 1$), and let $\{u_1, \ldots, u_m\}$ be any other frequency distribution satisfying the same conditions. Then using the fact that on the positive real line $\log(x) \leq (x - 1)$ with equality if and only if $x = 1$, we have

$$\sum_{j=1}^{m} f_j \log\left(\frac{u_j}{f_j}\right) \leq 0, \tag{9.72}$$

with equality if and only if $f_j = u_j$ for all j. In this we recognize the entropy expression (9.49), so the Gibbs inequality becomes

$$H(f_1, \ldots, f_m) \leq -\sum_{j=1}^{m} f_j \log(u_j), \tag{9.73}$$

from which various conclusions can be drawn. Making the choice $u_j = 1/m$ for all j, it becomes

$$H \leq \log(m), \tag{9.74}$$

so the maximum possible value of H is $\log(m)$, attained if and only if f_j is the uniform distribution $f_j = 1/m$ for all j. Now make the choice

$$u_j = \frac{\exp\left\{-\lambda g_j\right\}}{Z(\lambda)}, \tag{9.75}$$

where the normalizing factor $Z(\lambda)$ is just the partition function (9.21). Choose the constant λ so that some specified average $\overline{G} = \sum u_j g_j$ is attained; we shall see presently how to do this. The Gibbs inequality becomes

$$H \leq \sum f_j g_j + \log Z(\lambda). \tag{9.76}$$

Now let f_j vary over the class of all frequency distributions that yield the wanted average (9.71). The right-hand side of (9.76) remains constant, and H attains its maximum value on R:

$$H_{\max} = \overline{G} + \log(Z), \tag{9.77}$$

if and only if $f_j = u_j$. It remains only to choose λ so that the average value $\overline{G}$ is realized. But it is evident from (9.75) that

$$\overline{G} = -\frac{\partial}{\log(Z)\partial\lambda}, \tag{9.78}$$

so this is to be solved for λ. It is easy to see that this has only one real root (on the real axis, the right-hand side of (9.78) is a continuous, strictly decreasing monotonic function of λ), so the solution is unique.

We have just derived the 'Gibbs canonical ensemble' formalism, which in quantum statistics is able to determine all equilibrium thermodynamic properties of a closed system (that is, no particles enter it or leave it); but now its generality far beyond that application is evident.

9.10 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 James 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, whenever they are relevant to the problem, as mathematical consequences of probability theory as logic; 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.

R. A. Fisher, J. Neyman, R. von Mises, W. 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.

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 going back to James Bernoulli (1713), but it is also closer to the modern colloquial meaning of the word.

9.11 Significance tests

The rather subtle interplay between the notions of probability and frequency appears again in the topic of significance tests, or 'tests of goodness of fit'. In Chapter 5 we discussed such problems as assessing the validity of Newtonian celestial mechanics, and noted that orthodox significance tests purport to accept and reject hypotheses without considering any alternatives. Then we demonstrated why we cannot say how the observed facts affect the status of some hypothesis H until we state the specific alternative(s) against which H is to be tested. Common sense tells all scientists that a given piece of observational evidence E might demolish Newton's theory, or elevate it to certainty, or anything in between. It depends entirely on this: against which alternative(s) is it being tested? Bayes' theorem sends us the same message; for example, suppose we wish to consider only two hypotheses, H and H'. Then on any data D and prior information I, we must always have $P(H|DI) + P(H'|DI) = 1$, and in terms of our logarithmic measure of plausibility in decibels as discussed in Chapter 4, Bayes' theorem becomes

$$e(H|DI) = e(H|I) + 10\log_{10}\left[\frac{P(D|H)}{P(D|H')}\right], \tag{9.79}$$

which we might describe in words by saying that 'Data D supports hypothesis H relative to H' by $10\log_{10}[P(D|H)/P(D|H')]$ decibels'. The phrase '*relative to H'*' is essential here, because relative to some other alternative H'' the change in evidence, $[e(H|DI) - e(H|I)]$, might be entirely different; it does not make sense to ask how much the observed facts tend 'in themselves' to support or refute H (except, of course, when data D are impossible on hypothesis H, so deductive reasoning can take over).

Now as long as we talk only in these generalities, our common sense readily assents to this need for alternatives. But if we consider specific problems, we may have some doubts. For example, in the particle counter problem of Chapter 6 we had a case (known source strength and counter efficiency s, ϕ) where the probability for getting c counts in any one second is a Poisson distribution with mean value $\lambda = s\phi$:

$$p(c|s\phi) = \exp\{-\lambda\}\frac{\lambda^c}{c!}, \qquad 0 \le c \le \infty. \tag{9.80}$$

Although it wasn't necessary for the problem we were considering then, we can still ask: What can we infer from this about the relative *frequencies* with which we would see c counts if we repeat the measurement in many different seconds, with the resulting data set $D \equiv \{c_1, c_2, \ldots, c_n\}$? If the assigned probability for any particular event (say the event $c = 12$) is independently equal to

$$p = \exp\{-\lambda\}\frac{\lambda^{12}}{12!} \tag{9.81}$$

at each trial, then the probability that the event will occur exactly r times in n trials is the binomial distribution (9.38):

$$b(r|n, p) = \binom{n}{r} p^r (1-p)^{n-r}. \tag{9.82}$$

There are several ways of calculating the moments of this distribution; one, easy to remember, is that the first moment is

$$\begin{aligned}\langle r\rangle &= E(r)\\ &= \sum_{r=0}^{n} r b(r|n, p)\\ &= \left[p\frac{\mathrm{d}}{\mathrm{d}p}\sum_{r}\binom{n}{r}p^{r}q^{(n-r)}\right]_{q=1-p}\\ &= \left(p\frac{\mathrm{d}}{\mathrm{d}p}\right)\times(p+q)^{n}\\ &= np;\end{aligned} \tag{9.83}$$

likewise,

$$\begin{aligned}\langle r^2\rangle &= \left(p\frac{\mathrm{d}}{\mathrm{d}p}\right)^2(p+q)^n = np + n(n-1)p^2,\\ \langle r^3\rangle &= \left(p\frac{\mathrm{d}}{\mathrm{d}p}\right)^3(p+q)^n = np + 2n(n-1)p^2 + n(n-1)(n-2)p^3,\end{aligned} \tag{9.84}$$

and so on! For each higher moment we merely apply the operator $(p\mathrm{d}/\mathrm{d}p)$ one more time, setting $p + q = 1$ at the end.

Our (mean) ± (standard deviation) estimate, over the sampling distribution, of r is then

$$\begin{aligned}(r)_{\text{est}} &= \langle r\rangle \pm \sqrt{\langle r^2\rangle - \langle r\rangle^2}\\ &= np \pm \sqrt{np(1-p)},\end{aligned} \tag{9.85}$$

and our estimate of the frequency $f = r/n$ with which the event $c = 12$ will occur in n trials, is

$$(f)_{\text{est}} = p \pm \sqrt{\frac{p(1-p)}{n}}. \tag{9.86}$$

These relations and their generalizations give the most commonly encountered connection between probability and frequency; it is the original connection given by James Bernoulli (1713).

In the 'long run', therefore, we expect that the actual frequencies of various counts will be distributed in a manner approximating the Poisson distribution (9.80) to within the tolerances indicated by (9.86). Now we can perform the experiment, and the experimental frequencies either will or will not resemble the predicted values. If, by the time we have observed a few thousand counts, the observed frequencies are wildly different from a Poisson distribution (i.e. far outside the limits (9.86)), our intuition will tell us that the arguments which led to the Poisson prediction must be wrong; either the functional form of (9.80), or the independence at different trials, must not represent the real conditions in which the experiment was done.

Yet we have not said anything about any alternatives! Is our intuitive common sense wrong here, or is there some way we can reconcile it with probability theory? The question is not about probability theory but about psychology; it concerns what our intuition is doing here.

9.11.1 Implied alternatives

Let's look again at (9.79). No matter what H' is, we must have $p(D|H') \leq 1$, and therefore a statement which is independent of any alternative hypotheses is

$$e(H|DI) \geq e(H|I) + 10\log_{10} p(D|H) = e(H|I) - \psi_\infty, \tag{9.87}$$

where

$$\psi_\infty \equiv -10\log_{10} p(D|H) \geq 0. \tag{9.88}$$

Thus, *there is no possible alternative which data D could support, relative to H, by more than ψ_∞ decibels.*

This suggests the solution to our paradox: in judging the amount of agreement between theory and observation, the proper question to ask is not, 'How well do data D support hypothesis H?' without mentioning any alternatives. A much better question is, 'Are there any alternatives H' which data D would support relative to H, and how much support is possible?' Probability theory can give no meaningful answer to the first question because it is not well-posed; but it can give a very definite (quantitative and unambiguous) answer to the second.

We might be tempted to conclude that the proper criterion of 'goodness of fit' is simply ψ_∞; or what amounts to the same thing, just the probability $p(D|H)$. This is not so, however, as the following argument shows. As we noted at the end of Chapter 6, after we have obtained D, it is always possible to invent a strange, 'sure thing' hypothesis H_S according to which every detail of D was inevitable: $p(D|H_\mathrm{S}) = 1$, and H_S will always be supported relative to H by exactly ψ_∞ decibels. Let us see what this implies. Suppose we toss a die $n = 10\,000$ times and record the detailed results. Then, on the hypothesis $H \equiv$ 'the die is honest', each of the 6^n possible outcomes has probability 6^{-n}, or

$$\psi_\infty = 10\log_{10}(6^n) = 77\,815 \text{ db}. \tag{9.89}$$

No matter what we observe in the 10 000 tosses, there is always an hypothesis H_S that will be supported relative to H by this enormous amount. If, after observing 10 000 tosses, we still believe the die is honest, it can be only because we considered the prior probability for H_S to be even lower than $-77\,815$ db. Otherwise, we are reasoning inconsistently.

This is, if startling, all quite correct. The prior probability for H_S was indeed much lower than 6^{-n}, simply because there were 6^n different 'sure thing' hypotheses which were all on the same footing before we observed the data D. But it is obvious that in practice we don't want to bother with H_S; even though it is supported by the data more than any other, its prior probability is so low that we know in advance that we are not going to accept it anyway.

In practice, we are not interested in comparing H to all conceivable alternatives, but only to all those in some restricted class Ω, consisting of hypotheses which we consider in some sense 'reasonable'. Let us note one example (by far the most common and useful one) of a test relative to such a restricted class of alternatives.

Consider again the above experiment which has m possible results $\{A_1, \ldots, A_m\}$ at each trial. Define the quantities

$$x_i \equiv k, \qquad \text{if } A_k \text{ is true at the } i\text{th trial;} \tag{9.90}$$

thus, each x_i can take on independently the values $1, 2, \ldots, m$. Now we wish to take into account only the hypotheses belonging to the 'Bernoulli class' B_m in which there are m possible results at each trial and the probabilities of the A_k on successive repetitions of the experiment are considered independent and stationary; thus, when H is in B_m, the probability conditional on H of any specific sequence $\{x_1, \ldots, x_n\}$ of observations has the form

$$p(x_1 \ldots x_n|H) = p_1^{n_1} \ldots p_m^{n_m}, \tag{9.91}$$

where n_k is the above sample number. To every hypothesis in B_m there corresponds a set of numbers $\{p_1 \ldots p_m\}$ such that $p_k \geq 0$, $\sum_k p_k = 1$, and for our present purposes these numbers completely characterize the hypothesis. Conversely, every such set of numbers defines an hypothesis belonging to the Bernoulli class B_m.

Now we note an important lemma, given by J. Willard Gibbs (1902). Letting $x = n_k/np_k$, and using the fact that on the positive real line $\log(x) \geq (1 - x^{-1})$ with equality if and only if $x = 1$, we find at once that

$$\sum_{k=1}^{m} n_k \log\left(\frac{n_k}{np_k}\right) \geq 0, \tag{9.92}$$

with equality if and only if $p_k = n_k/n$ for all k. This inequality is the same as

$$\log p(x_1 \ldots x_n|H) \leq n \sum_{k=1}^{m} f_k \log(f_k), \tag{9.93}$$

where $f_k = n_k/n$ is the observed frequency of result A_k. The right-hand side of (9.88) depends only on the observed sample D, so if we consider various hypotheses $\{H_1, H_2, \ldots\}$ in B_m, the quantity (9.88) gives us a measure of how well the different hypotheses fit the data; the nearer to equality, the better the fit.

For convenience in numerical work, we express (9.88) in decibel units as in Chapter 4:

$$\psi_B \equiv 10 \sum_{k=1}^{m} n_k \log_{10}\left(\frac{n_k}{np_k}\right). \tag{9.94}$$

To see the meaning of ψ_B, suppose we apply Bayes' theorem in the form of (9.79). Only two hypotheses, $H = \{p_1, \dots, p_m\}$ and $H' = \{p'_1, \dots, p'_m\}$ are being considered. Let the values of ψ_B according to H and H' be ψ_B, ψ'_B, respectively. Then Bayes' theorem reads

$$\begin{aligned} e(H|x_1 \dots x_n) &= e(H|I) + 10\log_{10}\left[\frac{p(x_1 \dots x_n|H)}{p(x_1 \dots x_n|H')}\right] \\ &= e(H|I) + \psi'_B - \psi_B. \end{aligned} \tag{9.95}$$

Now we can always find an hypothesis H' in B_m for which $p'_k = n_k/n$, and so $\psi'_B = 0$; therefore ψ_B has the following meaning:

> Given an hypothesis H and the observed data $D \equiv \{x_1, \dots, x_n\}$, compute ψ_B from (9.94). Then, given any ψ in the range $0 \le \psi \le \psi_B$, it is possible to find an alternative hypothesis H' in B_m such that the data support H' relative to H by ψ decibels. There is no H' in B_m which is supported relative to H by more than ψ_B decibels.

Thus, although ψ_B makes no reference to any specific alternative, it is nevertheless exactly the appropriate measure of 'goodness of fit' *relative to the class B_m of Bernoulli alternatives*. It searches out B_m and locates the best alternative in that class.

Now we can understand the seeming paradox with which our discussion of significance tests started; the ψ-test is just the quantitative version of what our intuition has been, unconsciously, doing. We have already noted in Chapter 5, Section 5.4, that natural selection in exactly the sense of Darwin would tend to evolve creatures that reason in a Bayesian way because of its survival value.

We can also interpret ψ_B in this manner: we may regard the observed results $\{x_1, \dots, x_n\}$ as a 'message' consisting of n symbols chosen from an alphabet of m letters. On each repetition of the experiment, Nature transmits to us one more letter of the message. How much information is transmitted by this message under the Bernoulli probability assignment? Note that

$$\psi_B = 10n \sum_{k=1}^{m} f_k \log_{10}(f_k/p_k) \tag{9.96}$$

with $f_k = n_k/n$. Thus, $(-\psi_B/n)$ is the entropy per symbol $H(f; p)$ of the observed message distribution $\{f_1, \dots, f_m\}$ relative to the 'expected distribution' $\{p_1, \dots, p_m\}$. This shows that the notion of entropy was always inherent in probability theory; independently of Shannon's theorems, entropy or some monotonic function of entropy appears automatically in the equations of anyone who is willing to use Bayes' theorem for hypothesis testing.

Historically, a different criterion was introduced by Karl Pearson early in the 20th century. We expect that, if hypothesis H is true, then n_k will be close to np_k, in the sense that the difference $|n_k - np_k|$ will grow with n only as $\sqrt{n}$. Call this 'condition A'. Then using the expansion $\log(x) = (x-1) - (x-1)^2/2 + \cdots$, we find that

$$\sum_{k=1}^{m} n_k \log\left[\frac{n_k}{np_k}\right] = \frac{1}{2}\sum_k \frac{(n_k - np_k)^2}{np_k} + O\left(\frac{1}{\sqrt{n}}\right), \tag{9.97}$$

the quantity designated as $O(1/\sqrt{n})$ tending to zero as indicated *provided that* the observed sample does in fact satisfy condition A. The quantity

$$\chi^2 \equiv \sum_{k=1}^{m} \frac{(n_k - np_k)^2}{np_k} = n \sum_k \frac{(f_k - p_k)^2}{p_k} \tag{9.98}$$

is thus very nearly proportional to ψ_B if the sample frequencies are close to the expected values:

$$\psi_B = [10 \log_{10}(e)] \times \frac{1}{2}\chi^2 + O\left(\frac{1}{\sqrt{n}}\right) = 4.343\chi^2 + O\left(\frac{1}{\sqrt{n}}\right). \tag{9.99}$$

Pearson suggested that χ^2 be used as a criterion of 'goodness of fit', and this has led to the 'chi-squared test', one of the most used techniques of orthodox statistics. Before describing the test, we examine its theoretical basis and suitability as a criterion. Evidently, $\chi^2 \geq 0$, with equality if and only if the observed frequencies agree exactly with those expected if the hypothesis is true. So, larger values of χ^2 correspond to greater deviation between prediction and observation, and too large a value of χ^2 should lead us to doubt the truth of the hypothesis. But these qualitative properties are possessed also by ψ_B and by any number of other quantities we could define. We have seen how probability theory determines directly the theoretical basis, and precise quantitative meaning, of ψ_B; so we ask whether there exists any connected theoretical argument pointing to χ^2 as the optimal measure of goodness of fit, by some well-defined criterion.

The results of a search for this connected argument are disappointing. Scanning a number of orthodox textbooks, we find that χ^2 is usually introduced as a straight *deus ex machina*; but Cramér (1946) does attempt to prepare the way for the idea, in these words:

> It will then be in conformity with the general principle of least squares to adopt as measure of deviation an expression of the form $\sum c_i(n_i/n - p_i)^2$ where the coefficients c_i may be chosen more or less arbitrarily. It was shown by K. Pearson that if we take $c_i = n/p_i$, we shall obtain a deviation measure with particularly simple properties.

In other words, χ^2 is adopted, not because it is demonstrated to have good performance by any criterion, but only because it has simple properties!

We have seen that in some cases χ^2 is nearly a multiple of ψ_B, and then they must, of course, lead to essentially the same conclusions. But let us try to understand the quantitative difference in these criteria by a technique introduced in Jaynes (1976), which we borrowed from Galileo. Galileo's telescope was able to reveal the moons of Jupiter because it could *magnify* what was too small to be perceived by the unaided eye, up to the point where it could be seen by everybody. Likewise, we often find a quantitative difference in the Bayesian and orthodox results, so small that our common sense is unable to pass judgment on which result is preferable. But when this happens, we can find some extreme case where the difference is magnified to the point where common sense *can* tell us which method is giving sensible results, and which is not.

As an example of this magnification technique, we compare ψ_B and χ^2 to see which is the more reasonable *criterion* of goodness of fit.

9.12 Comparison of psi and chi-squared

A coin toss can give *three* different results: (1) heads, (2) tails, (3) it may stand on edge if it is sufficiently thick. Suppose that Mr A's knowledge of the thick English pound coin is such that he assigns probabilities $p_1 = p_2 = 0.499$, $p_3 = 0.002$ to these cases. We are in communication with Mr B on the planet Mars, who has never seen a coin and doesn't have the slightest idea what a coin is. So, when told that there are three possible results at each trial, and nothing more, he can only assign equal probabilities, $p'_1 = p'_2 = p'_3 = 1/3$.

Now we want to test Mr A's hypothesis against Mr B's by doing a 'random' experiment. We toss the coin 29 times and observe the results ($n_1 = n_2 = 14$, $n_3 = 1$). Then if we use the ψ criterion, we would have for the two hypotheses

$$\begin{aligned} \psi_A &= 10\left[28\log_{10}\left(\frac{14}{29\times 0.499}\right) + \log_{10}\left(\frac{1}{29\times 0.002}\right)\right] \\ &= 8.34\ \text{db}, \\ \psi_B &= 10\left[28\log_{10}\left(\frac{14\times 3}{29}\right) + \log_{10}\left(\frac{3}{29}\right)\right] \\ &= 35.19\ \text{db}. \end{aligned} \tag{9.100}$$

From this experiment, Mr B learns two things: (a) that there is another hypothesis about the coin that is 35.2 db better than his (this corresponds to odds of over 3 300:1), and so, unless he can justify an extremely low prior probability for that alternative, he cannot reasonably adhere to his first hypothesis; and (b) that Mr A's hypothesis is better than his by some 26.8 db, and in fact is within about 8 db of the best hypothesis in the Bernoulli class B_3. Here the ψ test tells us pretty much what our common sense does.

Suppose that the man on Mars knew only about orthodox statistical principles as usually taught; and therefore believed that χ^2 was the proper criterion of goodness of fit. He would find that

$$\begin{aligned} \chi^2_A &= 2\frac{(14 - 29\times 0.499)^2}{29\times 0.499} + \frac{(1 - 29\times 0.002)^2}{29\times 0.002} \\ &= 15.33, \\ \chi^2_B &= 2\frac{(14 - 29\times 0.333)^2}{29\times 0.333} + \frac{(1 - 29\times 0.333)^2}{29\times 0.333} \\ &= 11.66, \end{aligned} \tag{9.101}$$

and he would report back delightedly: 'My hypothesis, by the accepted statistical test, is shown to be slightly preferable to yours!'

Many persons trained to use χ^2 will find this comparison startling, and will try immediately to find the error in our numerical work above. We have here still another fulfilment of what Cox's theorems predict. The ψ criterion is exactly derivable from the rules of probability theory; therefore any criterion which is only an approximation to it must contain either an inconsistency or a qualitative violation of common sense, which can be exhibited by producing special cases.

We can learn an important lesson about the practical use of χ^2 by looking more closely at what is happening here. On hypothesis A, the 'expected' number of heads or tails in 29 tosses was $np_1 = 14.471$. The actual observed number must be an integer, and we supposed that in each case it was the closest possible integer, namely 14. Yet this small discrepancy between expected and observed sample numbers, in a sense the smallest it could possibly be, nevertheless had an enormous effect on χ^2. The spook lies in the fact that χ^2_A turned out so much larger than seems reasonable; there is nothing surprising about the other numerical values. Evidently, it is the last term in χ^2_A, which refers to the fact the coin stood on edge once in 29 tosses, that is causing the trouble. On hypothesis A, the probability that this would happen exactly r times in n tosses is our binomial distribution (9.57), and with $n = 29$, $p = 0.002$, we find that the probability for seeing the coin on edge one or more times in 29 trials is $1 - b(0|n, p) = 1 - 0.998^{29} = 1/17.73$; i.e. the fact that we saw it even once is a bit unexpected, and constitutes some evidence against A. But this amount of evidence is certainly not overwhelming; if our travel guide tells us that London has fog, on the average, one day in 18, we are hardly astonished to see fog on the day we arrive. Yet this contributes an amount 15.30, almost all of the value of $\chi^2_A = 15.33$.

It is the $(1/p_i)$ weighting factor in the summand of χ^2 that causes this anomaly. Because of it, the χ^2 criterion essentially concentrates its attention on the extremely unlikely possibilities if the hypothesis contains them; and the slightest discrepancy between expected and observed sample number for the unlikely events grotesquely over-penalizes the hypothesis. The ψ-test also contains this effect, but in a much milder form, the $1/p_i$ term appearing only in the logarithm.

To see this effect more clearly, suppose now that the experiment had yielded instead the results $n_1 = 14, n_2 = 15, n_3 = 0$. Evidently, by either the χ^2 or ψ criterion, this ought to make hypothesis A look better, B worse, than in the first example. Repeating the calculations, we now find

$$\begin{aligned} \psi_A &= 0.30 \text{ db} \qquad & \chi^2_A &= 0.0925 \\ \psi_B &= 51.2 \text{ db} \qquad & \chi^2_B &= 14.55. \end{aligned} \tag{9.102}$$

You see that by far the greatest relative change was in χ^2_A; both criteria now agree that hypothesis A is far superior to B, as far as this experiment indicates.

This shows what can happen through uncritical use of χ^2. Professor Q believes in extrasensory perception, and undertakes to prove it to us poor benighted, intransigent doubters. So he plays card games. As in Chapter 5, on the 'null hypothesis' that only chance is operating, it is extremely unlikely that the subject will guess many cards correctly. But Professor Q is determined to avoid the tactical errors of his predecessors, and is alert to the phenomenon of deception hypotheses discussed in Chapter 5; so he averts that possibility by making videotape recordings of every detail of the experiments. The first few hundred times he plays, the results are disappointing; but these are readily explained away on the grounds that the subject is not in a 'receptive' mood. Of course, the tapes recording these experiments are erased.

One day, providence smiles on Professor Q; the subject comes through handsomely and he has the incontrovertible record of it. Immediately he calls in the statisticians, the mathematicians, the notary publics, and the newspaper reporters. An extremely improbable event has at last occurred; and χ^2 is enormous. Now he can publish the results and assert: 'The validity of the data is certified by reputable, disinterested persons, the statistical analysis has been under the supervision of recognized statisticians, the calculations have been checked by competent mathematicians. By the accepted statistical test, the null hypothesis has been decisively rejected.' And everything he has said is absolutely true!

Moral

For testing hypotheses involving moderately large probabilities, which agree moderately well with observation, it will not make much difference whether we use ψ or χ^2. But for testing hypotheses involving extremely unlikely events, we had better use ψ; or life might become too exciting for us.

9.13 The chi-squared test

Now we examine briefly the chi-squared test as done in practice. We have the so-called 'null hypothesis' H to be tested, and no alternative is stated. The null hypothesis predicts certain relative frequencies $\{f_1, \dots, f_m\}$ and corresponding sample numbers $n_k = nf_k$, where n is the number of trials. We observe the actual sample numbers $\{n_1, \dots, n_m\}$. But if the n_k are very small, we group categories together, so that each n_k is at least, say, five. For example, in a case with $m = 6$, if the observed sample numbers were {6, 11, 14, 7, 3, 2} we would group the last two categories together, making it a problem with $m = 5$ distinguishable results per trial, with sample numbers {6, 11, 14, 7, 5}, and null hypothesis H which assigns probabilities $\{p_1, p_2, p_3, p_4, p_5 + p_6\}$. We then calculate the observed value of χ^2:

$$\chi^2_{\text{obs}} = \sum_{k=1}^{m} \frac{(n_k - np_k)^2}{np_k} \tag{9.103}$$

as our measure of deviation of observation from prediction. Evidently, it is very unlikely that we would find $\chi^2_{\text{obs}} = 0$ even if the null hypothesis is true. So, goes the orthodox reasoning, we should calculate the probability that χ^2 would have various values, and reject H if the probability $P(\chi^2_{\text{obs}})$ of finding a deviation as great as *or greater than* χ^2_{obs} is sufficiently small; this is the 'tail area' criterion, and one usually takes 5% (that is, $P(\chi^2_{\text{obs}}) = 0.05$) as the threshold of rejection.

Now the n_k are integers, so χ^2 is capable of taking on only a discrete set of numerical values, at most $(n + m - 1)!/n!(m - 1)!$ different values if the p_k are all different and incommensurable. Therefore, the exact χ^2 distribution is necessarily discrete and defined at only a finite number of points. However, for sufficiently large n, the number and density of points becomes so large that we may approximate the true χ^2 distribution by a continuous one. The ‘simple property’ referred to by Cramér is then the fact, at first glance surprising, that, in the limit of large n, we obtain a *universal* distribution law: the sampling probability that χ^2 will lie in the interval $\mathrm{d}(\chi^2)$ is

$$g(\chi^2)\mathrm{d}(\chi^2) = \frac{\chi^{f-2}}{2^{f/2}(f/2-1)!}\exp\left\{-\frac{1}{2}\chi^2\right\}\mathrm{d}(\chi^2), \tag{9.104}$$

where f is called the ‘number of degrees of freedom’ of the distribution. If the null hypothesis H is completely specified (i.e., if it contains no variable parameters), then $f = m - 1$, where m is the number of categories used in the sum of (9.98). But if H contains unspecified parameters which must be estimated from the data, we take $f = m - 1 - r$, where r is the number of parameters estimated.[5]

We readily calculate the expectation and variance over this distribution: $\langle\chi^2\rangle = f$, $\mathrm{var}(\chi^2) = 2f$, so the (mean) ± (standard deviation) estimate of the χ^2 that we expect to see is just

$$\left(\chi^2\right)_{\text{est}} = f \pm \sqrt{2f}. \tag{9.105}$$

The reason usually given for grouping categories for which the sample numbers are small, is that the approximation (9.104) would otherwise be bad. But grouping inevitably throws away some of the relevant information in the data, and there is never any reason to do this when using the exact ψ.

The probability that we would see a deviation as great as or greater than χ^2_{obs} is then

$$P(\chi^2_{\text{obs}}) = \int_{\chi^2_{\text{obs}}}^{\infty}\mathrm{d}(\chi^2)\,g(\chi^2) = \int_{q_{\text{obs}}}^{\infty}\mathrm{d}q\,\frac{q^k}{k!}\exp\{-q\}, \tag{9.106}$$

where $q \equiv (1/2)\chi^2$, $k \equiv (f-2)/2$. If $P(\chi^2_{\text{obs}}) < 0.05$, we reject the null hypothesis at the 5% ‘significance level’. Tables of χ^2 for which $P = 0.01, 0.05, 0.10, 0.50$, for various numbers of degrees of freedom, are given in most orthodox textbooks and collections of statistical tables (for example, Crow, Davis, and Maxfield, 1960).

Note the traditional procedure here; we chose some basically arbitrary significance level, then reported only whether the null hypothesis was or was not rejected at that level. Evidently this doesn’t tell us very much about the real import of the data; if you tell me that the hypothesis was rejected at the 5% level, then I can’t tell from that whether it would have been rejected at the 1%, or 2%, level. If you tell me that it was not rejected at the 5% level, then I don’t know whether it would have been rejected at the 10%, or 20%, level. The orthodox statistician would tell us far more about what the data really indicate if he would report instead *the significance level* $P(\chi^2)$ *at which the null hypothesis is just barely rejected*; for then we know what the verdict would be at all levels. This is the practice of reporting

[5] The need for this correction was perceived by the young R. A. Fisher but not comprehended by Karl Pearson; and this set off the first of their fierce controversies, described in Chapter 16.

so-called 'P-values', a major improvement over the original custom. Unfortunately, the orthodox χ^2 and other tables are still so constructed that you cannot use them to report the conclusions in this more informative way, because they give numerical values only at such widely separated values of the significance level that interpolation is not possible.

How does one find numerical P-values without using the chi-squared tables? Writing $q = q_0 + t$, (9.106) becomes

$$\begin{aligned} P &= \int_0^\infty \mathrm{d}t \, \frac{(q_0+t)^k}{k!} \exp\{-(q_0+t)\} \\ &= \frac{1}{k!} \sum_{k=0}^{m} \binom{m}{k} \int_0^\infty \mathrm{d}t \, q_0^k t^{m-k} \exp\{-(q_0+t)\} \\ &= \sum_{k=0}^{m} \exp\{-q_0\} \frac{q_0^k}{k!}. \end{aligned} \tag{9.107}$$

But this is just the cumulative Poisson distribution and easily computed.

If you use the ψ-test instead, however, you don't need any tables. The evidential meaning of the sample is then described simply by the *numerical value* of ψ, and not by a further arbitrary construct such as tail areas. Of course, the numerical value of ψ doesn't in itself tell you whether to reject the hypothesis (although we could, with just as much justification as in the chi-squared test, prescribe some definite 'level' at which to reject). From the Bayesian point of view, there is simply no use in rejecting any hypothesis unless we can replace it with a definite alternative known to be better; and, obviously, whether this is justified must depend not only on ψ, but also on the prior probability for the alternative and on the consequences of making wrong decisions. Common sense tells us that this is, necessarily, a problem not just of inference, but of decision theory.

In spite of the vast difference in viewpoints, there is not necessarily much difference in the actual conclusions reached. For example, as the number of degrees of freedom f increases, the orthodox statistician will accept a higher value of χ^2 (roughly proportional to f, as (9.105) indicates) on the grounds that such a high value is quite likely to occur if the hypothesis is true; but the Bayesian, who will reject it only in favor of a definite alternative, must also accept a proportionally higher value of ψ, because the number of reasonable alternatives is increasing exponentially with f, and the prior probability for any one of them is correspondingly decreasing. So, in either case we end up rejecting the hypothesis if ψ or χ^2 exceeds some critical limit, with an enormous difference in the philosophy of how we choose that limit, but not necessarily a big difference in its actual location.

For many more details about chi-squared, see Lancaster (1969); and for some curious views that Bayesian methods fail to give proper significance tests, see Box and Tiao (1973).

9.14 Generalization

Although the point is not made in the orthodox literature, which does not mention alternatives at all, we see from the preceding section that χ^2 is not a measure of goodness of fit relative to all conceivable alternatives, but only relative to those in the same Bernoulli class. Until this is recognized, one really does not know what the χ^2-test is testing.

The procedure by which we constructed the ψ-test generalizes at once to the rule for constructing the exact test which compares the null hypothesis to any well-defined class C of alternatives. Just write Bayes' theorem describing the effect of data D on the relative plausibility of two hypotheses H_1, H_2 in that class, in the form

$$e(H_1|DI) - e(H_2|DI) = \psi_2 - \psi_1, \tag{9.108}$$

where ψ_i depends only on the data and H_i is non-negative over C, and vanishes for some H_i in C. Then we can always find an H_2 in C for which $\psi_2 = 0$, and so we have constructed the appropriate ψ_1 which measures goodness of fit relative to the class of alternatives C, and has the same meaning as that defined after (9.95). ψ_1 is the maximum amount by which any hypothesis in C can be supported relative to H_1 by the data D.

Thus, if we want a Bayesian test that is exact but operates in a similar way to orthodox significance tests, it can be produced quite easily. But we shall see in Chapter 17 that a different viewpoint has advantages; the format of orthodox significance tests can be replaced, as was done already by Laplace, by a parameter estimation procedure, which yields even more useful information.

Anscombe (1963) held it to be a weakness of the Bayesian method that we had to introduce a specific class of alternatives. We have answered that sufficiently here and in Chapters 4 and 5. We would hold it to be a great merit of the Bayesian approach that it forces us to recognize these essential features of inference, which have not been apparent to all orthodox statisticians. Our discussion of significance tests is a good example of what, we suggest, is the general situation; if an orthodox method is usable in some problem, then the Bayesian approach to inference supplies the missing theoretical basis for it, and usually improvements on it. Any significance test is only a slight variant of our multiple hypothesis testing procedures given in Chapter 4.

9.15 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 16th century up to the adoption of public sanitation policies and facilities in the mid-19th 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 'ancient

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 (1620–74) 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 (Graunt, 1662).[6] 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.[7]

Edmund Halley was highly educated, mathematically competent (later succeeding Wallis (in 1703) as Savilian Professor of Mathematics at Oxford University and Flamsteed (in 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. Those data were so contaminated with trend that he had no means of extracting the information he needed. So he found instead five years' data (1687–91) 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 Figure 9.1.

In the 17th 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

[6] 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 1% 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.

[7] 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: against what alternative do we judge him? 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 19th century would not have happened. It is even problematical whether Cambridge and Oxford Universities would still exist today.

Table 9.1. *Halley's first table.*

Age	$d(y)/5$	Age	$d(y)/5$	Age	$d(y)/5$	Age	$d(y)/5$
0	348	28	8	⋮	10	90	1
⋮	198	⋮	7	63	12	91	1
7	11	35	7	⋮	9.5	98	0
8	11	36	8	70	14	99	0.5
9	6	⋮	9.5	71	9	100	3/5
⋮	5.5	42	8	72	11		
14	2	⋮	9	⋮	9.5		
⋮	3.5	45	7	77	6		
18	5	⋮	7	⋮	7		
⋮	6	49	10	81	3		
21	4.5	54	11	⋮	4		
⋮	6.5	55	9	84	2		
27	9	56	9	⋮	1		

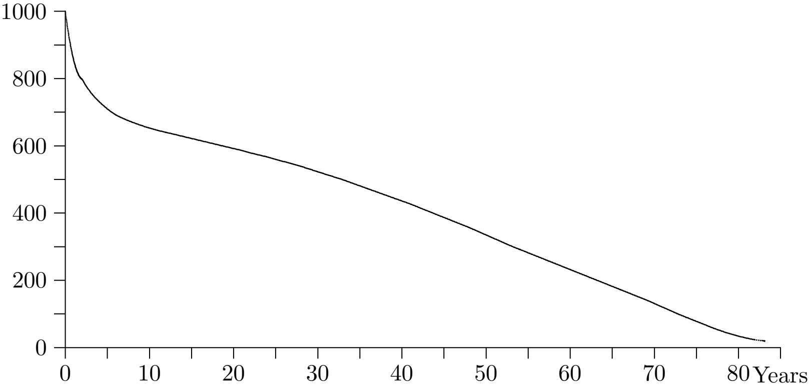

Fig. 9.1. $n(y)$: estimated number of persons in the age range $(y, y+1)$ years.

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, Table 9.1 (but which inexplicably contains some entries that are not multiples of 1/5), and the estimated distribution $n(x)$ of population by age, Table 9.2. Thus, the first table is, crudely, something like

Table 9.2. *Halley's second table.*

Age	$n(y)$	Age	$n(y)$	Age	$n(y)$	Age	$n(y)$	Age	$n(y)$	Age	$n(y)$	Age	$n(y)$
1	1000	13	640	25	567	37	472	49	357	61	232	73	109
2	855	14	634	26	560	38	463	50	346	62	222	74	98
3	798	15	628	27	553	39	454	51	335	63	212	75	88
4	760	16	622	28	546	40	445	52	324	64	202	76	78
5	732	17	616	29	539	41	436	53	313	65	192	77	68
6	710	18	610	30	531	42	427	54	302	66	182	78	58
7	692	19	604	31	523	43	417	55	292	67	172	79	49
8	680	20	598	32	515	44	407	56	282	68	162	80	41
9	670	21	592	33	507	45	397	57	272	69	152	81	34
10	661	22	586	34	499	46	387	58	262	70	142	82	28
11	653	23	579	35	490	47	377	59	252	71	131	83	23
12	646	24	573	36	481	48	367	60	242	72	120	84	20

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 Halley 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. Firstly, 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 the 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 9.2 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 9.2.

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 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. At this point we must give up, and simply accept Halley’s judgment, whatever it was.

In Figure 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 $n(84), \dots, n(99)$), 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 9.2) 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 a 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 far 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?

9.16 Comments

9.16.1 The irrationalists

Philosophers have argued over the nature of induction for centuries. Some, from David Hume (1711–76) in the mid-18th century to Karl Popper in the mid-20th (for example, Popper and Miller, 1983), have tried to deny the possibility of induction, although all scientific knowledge has been obtained by induction. D. Stove (1982) calls them and their colleagues '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, we are not bothered by this situation because we are not convinced that much of an audience exists.

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, Newton's theory in Chapter 5, and the above discussion of significance tests, 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 in order 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. As Harold Jeffreys explained long ago, induction is most valuable to a scientist just when it turns out to be wrong; only then do we get new fundamental knowledge.

Some striking case histories of induction in use 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 20th century that new influenza pandemics were coming out of China; the worst ones acquired names like the Asian Flu (in 1957), the Hong Kong Flu (in 1968), and Beijing A (in 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 unraveling 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 Hume and Popper held to be impossible. It is true that this reasoning format does not guarantee that the conclusion *must* be correct; but then direct tests can confirm it or refute it. Without the preparatory inductive reasoning phase, one would not know which direct tests to try.

9.16.2 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.[8]

Thus, the commentators for the very numerous television 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. Every 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.[9]

It would seem that we have here a valuable opportunity to illustrate and explain evolution; yet the commentators (usually out-of-work actors) have no comprehension of the simple, easily understood cause-and-effect mechanism pointed out over 100 years ago 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 unlikely that a paleontologist will ever find evidence of them.[10] 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

[8] 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.

[9] But it is hard to believe that the ridiculous color patterns of 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.

[10] 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.

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. Indeed, the creatures arriving at any environmental niche are seldom perfectly adapted to it; often they are just barely well enough adapted to survive. But then, in this stressful situation, bad mutations are eliminated faster than usual, so natural selection operates faster than usual to make them better adapted.