CHAPTER 3

ELEMENTARY SAMPLING THEORY

At this point, the mathematical material we have available consists of the basic product and sum rules

$$P(AB|C) = P(A|BC)\,P(B|C) = P(B|AC)\,P(A|C) \tag{3–1}$$

$$P(A|B) + P(\overline{A}|B) = 1 \tag{3–2}$$

from which we derived the extended sum rule

$$P(A+B|C) = P(A|C) + P(B|C) - P(AB|C) \tag{3–3}$$

and with the desideratum (III.c) of consistency, the principle of indifference: if on background information B the hypotheses $(H_1 \cdots H_N)$ are mutually exclusive and exhaustive, and B does not favor any one of them over any other, then

$$P(H_i|B) = \frac{1}{N}\,, \qquad 1 \le i \le N\,. \tag{3–4}$$

From (3–3) and (3–4) we then derived the Bernoulli urn rule; if B specifies that A is true on some subset of M of the H_i, false on the remaining $(N - M)$, then

$$P(A|B) = \frac{M}{N}\,. \tag{3–5}$$

It is important to realize how much of probability theory can be derived from no more than this.

In fact, essentially all of conventional probability theory as currently taught, plus many important results that are often thought to lie beyond the domain of probability theory, can be derived from the above foundation. We devote the next several Chapters to demonstrating this in some detail, and then in Chapter 11 resume the basic development of our robot's brain, with a better understanding of what additional principles are needed for advanced applications.

The first applications of the theory given in this Chapter are, to be sure, rather simple and naïve compared to the serious scientific inference that we hope to achieve later. Nevertheless, our reason for considering them in close detail is not mere pedagogical form. Failure to understand the logic of these simplest applications has been one of the major factors retarding the progress of scientific inference – and therefore of science itself – for many decades. Therefore we urge the reader, even one who considers himself already familiar with elementary sampling theory, to digest the contents of this Chapter carefully before proceeding to more complicated problems.

Sampling Without Replacement.

Let us make the Bernoulli Urn scenario a little more specific by defining the propositions:

$B \equiv$ "An urn contains N balls, identical in every respect except that they carry numbers $(1, 2 \ldots N)$ and M of them are colored red, the remaining $(N - M)$ white, $0 \le M \le N$. We draw a ball from the urn blindfolded, observe and record its color, lay it aside, and repeat the process until n balls have been drawn, $0 \le n \le N$."

$R_i \equiv$ "Red ball on the i'th draw."

$W_i \equiv$ "White ball on the i'th draw," Since according to B only red or white can be drawn, we have

$$P(R_i|B) \ + \ P(W_i|B) \ = \ 1 \ , \qquad 1 \le i \le N \tag{3–6}$$

which amounts to saying that, in the "logical environment" created by knowledge of B, the propositions are related by negation:

$$\overline{R}_i \ = \ W_i \ , \qquad \overline{W}_i \ = \ R_i \tag{3–7}$$

and for the first draw, (3–5) becomes

$$P(R_1|B) = \frac{M}{N} \ , \tag{3–8}$$

$$P(W_1|B) = 1 - \frac{M}{N} \ , \tag{3–9}$$

Let us understand clearly what this means. The probability assignments (3–8), (3–9) are not assertions of any physical property of the urn or its contents; they are a description of the *state of knowledge* of the robot prior to the drawing. Indeed, were the robot's state of knowledge different from B as just defined (for example, if it knew the actual positions of the red and white balls in the urn, or if it did not know the true values of N and M), then its probability assignments for R_1 and W_1 would be different; but the real properties of the urn would be just the same.

It is therefore illogical to speak of "verifying" (3–8) by performing experiments with the urn; that would be like trying to verify a boy's love for his dog by performing experiments on the dog. At this stage, we are concerned with the logic of consistent reasoning from incomplete information; not with assertions of physical fact about what will be drawn from the urn (which are in any event impossible just because of the incompleteness of the information B).

Eventually, our robot will be able to make some very confident physical predictions which can approach, but (except in degenerate cases) not actually reach, the certainty of logical deduction; but the theory needs to be developed further before we are in a position to say what quantities can be well predicted, and what kind of information is needed for this. Put differently, relations between probabilities assigned by the robot in various states of knowledge, and observable facts in experiments, may not be assumed arbitrarily; we are justified in using only those relations that can be deduced from the rules of probability theory, as we now seek to do.

Changes in the robot's state of knowledge appear already when we ask for probabilities referring to the second draw. For example, what is the robot's probability for red on the first two draws? From the product rule, this is

$$P(R_1 R_2|B) \ = \ P(R_1|B) \ P(R_2|R_1 B) \ .$$

In the last factor, the robot must take into account that one red ball has been removed at the first draw, so there remain $(N-1)$ balls of which $(M-1)$ are red. Therefore

$$P(R_1 R_2|B) \ = \ \frac{M}{N} \, \frac{M-1}{N-1} \ . \tag{3–10}$$

Continuing in this way, the probability of red on the first r consecutive draws is

$$\begin{aligned} P(R_1 \cdots R_r|B) \ &= \ \frac{M(M-1)\cdots(M-r+1)}{N(N-1)\cdots(N-r+1)} \\ &= \frac{M!(N-r)!}{(M-r)!N!} \ , \qquad r \le M \end{aligned} \tag{3–11}$$

The restriction $r \le M$ is not necessary if we understand that we define factorials by the gamma function relation $n! = \Gamma(n+1)$, for then the factorial of a negative integer is infinite, and (3–11) is zero automatically when $r > M$.

The probability of white on the first w draws is similar but for the interchange of M and $(N-M)$:

$$P(W_1 \;...\; W_w|B) = \frac{(N-M)!}{(N-M-w)!}\,\frac{(N-w)!}{N!} \tag{3–12}$$

Then, the probability of white on draws $(r+1,\, r+2, \ldots,\, r+w)$ given that we got red on the first r draws, is given by (3–12) taking into account that N and M have been reduced to $(N-r)$ and $(M-r)$:

$$P(W_{r+1} \ldots W_{r+w}|R_1 \ldots R_r B) \;=\; \frac{(N-M)!\,(N-r-w)!}{(N-M-w)!\,(N-r)!} \tag{3–13}$$

and so, by the product rule, the probability of obtaining r red followed by $w = n-r$ white in n draws is from (3–11), (3–13),

$$P(R_1 \cdots R_r\; W_{r+1} \ldots W_n|B) = \frac{M!\,(N-M)!\,(N-n)!}{(M-r)!\,(N-M-w)!\;N!} \tag{3–14}$$

a term $(N-r)!$ having cancelled out.

Although this result was derived for a particular order of drawing red and white balls, the probability of drawing exactly r red balls in any specified order in n draws is the same. To see this, write out the expression (3–14) more fully, in the manner

$$\frac{M!}{(M-r)!} \;=\; M(M-1)\cdots(M-r+1)$$

and similarly for the other ratios of factorials in (3–14). The right-hand side becomes

$$\frac{M(M-1)\cdots(M-r+1)\;(N-M)(N-M-1)\cdots(N-M-w+1)}{N(N-1)\cdots(N-n+1)}\;. \tag{3–15}$$

Now suppose that r red and $(n-r) = w$ white are drawn, in any other order. The probability of this is the product of n factors; every time red is drawn there is a factor (number of red balls in urn)/(total number of balls), and similarly for drawing a white one. The number of balls in the urn decreases by one at each draw; therefore for the k'th draw a factor $(N-k+1)$ appears in the denominator, whatever the colors of the previous draws.

Just before the k'th red ball is drawn, whether this occurs at the k'th draw or any later one, there are $(M-k+1)$ red balls in the urn; so drawing the k'th one places a factor $(M-k+1)$ in the numerator. Just before the k'th white ball is drawn, there are $(N-M-k+1)$ white balls in the urn, and so drawing the k'th white one places a factor $(N-M-k+1)$ in the numerator, regardless of whether this occurs at the k'th draw or any later one. Therefore, by the time all n balls have been drawn, of which r were red, we have accumulated exactly the same factors in numerator and denominator as in (3–15); different orders of drawing them only permute the order of the factors in the numerator. The probability of drawing exactly r balls in any specified order in n draws, is therefore given by (3–14).

Note carefully that in this result the product rule was expanded in a particular way that showed us how to organize the calculation into a product of factors, each of which is a probability at one specified draw, *given the results of all the previous draws.* But the product rule could have been expanded in many other ways, which would give factors conditional on other information than the previous draws; the fact that all these calculations must lead to the same final result is a nontrivial consistency property, which the derivations of Chapter 2 sought to ensure.

Next, we ask: what is the robot's probability for drawing exactly r red balls in n draws, regardless of order? Different orders of appearance of red and white balls are mutually exclusive possibilities, so we must sum over all of them; but since each term is equal to (3–14), we merely multiply it by the binomial coefficient

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!} \tag{3–16}$$

which represents the number of possible orders of drawing r red balls in n draws or as we shall call it, the *multiplicity* of the event r. For example, to get 3 red in 3 draws can happen in only

$$\binom{3}{3} = 1$$

way, namely $R_1 R_2 R_3$; the event $r = 3$ has a multiplicity of 1. But to get 2 red in 3 draws can happen in

$$\binom{3}{2} = 3$$

ways, namely $R_1 R_2 W_3$, $R_1 W_2 R_3$, $W_1 R_2 R_3$, so the event $r = 2$ has a multiplicity of 3.

Exercise 3.1. Why isn't the multiplicity factor (3–16) just $n!$? After all, we started this discussion by stipulating that the balls, in addition to having colors, also carry labels $(1 \cdots N)$, so that different permutations of the red balls among themselves, which give the $r!$ in the denominator of (3–16), are distinguishable arrangements. *Hint:* in (3–14) we are not specifying which red balls and which white ones are to be drawn.

Then taking the product of (3–14) and (3–16), the many factorials can be reorganized into three binomial coefficients. Defining $A \equiv$ "Exactly r red balls in n draws, in any order" and the function

$$h(r|N,M,n) \equiv P(A|B) \tag{3–17}$$

we have

$$h(r|N,M,n) = \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}} \tag{3–18}$$

which we shall usually abbreviate to $h(r)$. By the convention $x! = \Gamma(x+1)$ it vanishes automatically when $r > M$, or $r > n$, or $(n-r) > (N-M)$, as it should.

We are here doing a little notational acrobatics for reasons explained in Appendix B. The point is that in our formal probability symbols $P(A|B)$ with the capital P, the arguments A, B always stand for propositions, which can be quite complicated verbal statements. If we wish to use ordinary numbers for arguments, then for consistency we should define new functional symbols such as $h(r|N,M,n)$. To try to use a notation like $P(r|N,M,n)$, thereby losing sight of the qualitative stipulations contained in A and B, has led to serious errors from misinterpretation of the equations (such as the marginalization paradox discussed later). However, as already indicated in Chapter 2, we follow the custom of most contemporary works by using probability symbols of the form $p(A|B)$, or $p(r|n)$ with small p, in which we permit the arguments to be either propositions or algebraic variables; in this case, the meaning must be judged from the context.

The fundamental result (3–18) is called the *hypergeometric distribution* because it is related to the coefficients in the power series representation of the Gauss hypergeometric function

$$F(a,b,c;t) \;=\; \sum_{r=0}^{\infty} \frac{\Gamma(a+r)\Gamma(b+r)\Gamma(c)}{\Gamma(a)\Gamma(b)\Gamma(c+r)}\,\frac{t^r}{r!}\,. \tag{3–19}$$

If either a or b is a negative integer, the series terminates and this is a polynomial. It is easily verified that the *generating function*

$$G(t) \equiv \sum_{r=0}^{n} h(r|N,M,n)\,t^r \tag{3–20}$$

is equal to

$$G(t) \;=\; \frac{F(-M,-n,c;t)}{F(-M,-n,c;1)} \tag{3–21}$$

with $c = N - M - n + 1$. The evident relation $G(1) = 1$ is from (3–20) just the statement that the hypergeometric distribution is correctly normalized. In consequence of (3–21), $G(t)$ satisfies the second–order hypergeometric differential equation and has many other properties useful in calculations. Further details about generating functions are in Appendix D.

Although the hypergeometric distribution $h(r)$ appears complicated, it has some surprisingly simple properties. The most probable value of r is found to within one unit by setting $h(r') = h(r'-1)$ and solving for r'. We find

$$r' = \frac{(n+1)(M+1)}{N+2}\,. \tag{3–22}$$

If r' is an integer, then r' and $r'-1$ are jointly the most probable values. If r' is not an integer, then there is a unique most probable value

$$\hat{r} = INT(r') \tag{3–23}$$

that is, the next integer below r'. Thus the most probable fraction $f = r/n$ of red balls in the sample drawn is nearly equal to the fraction $F = M/N$ originally in the urn, as one would expect intuitively. This is our first crude example of a physical prediction: a relation between a quantity F specified in our information, and a quantity f measurable in a physical experiment, derived from the theory.

The width of the distribution $h(r)$ gives an indication of the accuracy with which the robot can predict r. Many such questions are answered by calculating the *cumulative probability distribution*, which is the probability of finding R or fewer red balls. If R is an integer, that is

$$H(R) \equiv \sum_{r=0}^{R} h(r) \quad , \tag{3–24}$$

but for later formal reasons we define $H(x)$ to be a staircase function for all real x; thus $H(x) \equiv H(R)$, where $R = \mathrm{INT}(x)$ is the greatest integer $\leq x$.

The *median* of a probability distribution such as $h(r)$ is defined to be a number m such that equal probabilities are assigned to the propositions $(r < m)$ and $(r > m)$. Strictly speaking, according to this definition a discrete distribution has in general no median. If there is an integer

R for which $H(R-1) = 1 - H(R)$ and $H(R) > H(R-1)$, then R is the unique median. If there is an integer R for which $H(R) = 1/2$, then any r in $(R \le r < R')$ is a median, where R' is the next higher jump point of $H(x)$; otherwise there is none.

But for most purposes we may take a more relaxed attitude and approximate the strict definition. If n is reasonably large, then it makes reasonably good sense to call that value of R for which $H(R)$ is closest to $1/2$, the "median". In the same relaxed spirit, the values of R for which $H(R)$ is closest to $1/4$, $3/4$ may be called the "lower quartile" and "upper quartile", and if $n >> 10$ we may call the value of R for which $H(R)$ is closest to $k/10$ the "k'th decile", and so on. As $n \to \infty$ these loose definitions come into conformity with the strict one.

Usually, the fine details of $H(R)$ are unimportant and for our purposes it is sufficient to know the median and the quartiles. Then the (median) $\pm$ (interquartile distance) will provide a good enough idea of the robot's prediction and its probable accuracy. That is, on the information given to the robot, the true value of r is about as likely to lie in this interval as outside it. Likewise, the robot assigns a probability of $(5/6) - (1/6) = 2/3$ (in other words, odds of $2:1$) that r lies between the first and fifth hexile, odds of $8:2 = 4:1$ that it is bracketed by the first and ninth decile; and so on.

Although one can develop rather messy approximate formulas for these distributions which were much used in the past, it is easier today to calculate the exact distribution by computer. In Appendix I we give a short program HYPERGEO.BAS which will run on almost any microcomputer, and which prints out $h(r)$ and $H(R)$ for N up to 130. Beyond that, the binomial approximation given below will be accurate enough.

For example, Tables 3.1 and 3.2 give the HYPERGEO printouts for $N = 100$, $M = 50$, $n = 10$ and $N = 100$, $M = 10$, $n = 50$. In the latter case, it is not possible to draw more than 10 red balls, so the entries for $r > 10$ are all $h(r) = 0$, $H(r) = 1$ and are not tabulated. One is struck immediately by the fact that the entries for positive $h(r)$ are identical; the hypergeometric distribution has the symmetry property

$$h(r|N, M, n) = h(r|N, n, M) \tag{3–25}$$

under interchange of M and n. Whether we draw 10 balls from an urn containing 50 red ones, or 50 from an urn containing 10 red ones, the probability of finding r red ones in the sample drawn is the same. This is readily verified by closer inspection of (3–18), and it is evident from the symmetry in a, b of the hypergeometric function (3–19).

r	$h(r)$	$H(r)$
0	0.000593	0.000593
1	0.007237	0.007830
2	0.037993	0.045824
3	0.113096	0.158920
4	0.211413	0.370333
5	0.259334	0.629667
6	0.211413	0.841080
7	0.113096	0.954177
8	0.037993	0.992170
9	0.007237	0.999407
10	0.000593	1.000000

Table 3.1: $N, M, n = 100, 10, 50.$

r	$h(r)$	$H(r)$
0	0.000593	0.000593
1	0.007237	0.007830
2	0.037993	0.045824
3	0.113096	0.158920
4	0.211413	0.370333
5	0.259334	0.629667
6	0.211413	0.841080
7	0.113096	0.954177
8	0.037993	0.992170
9	0.007237	0.999407
10	0.000593	1.000000

Table 3.2: $N, M, n = 100, 50, 10$

Another symmetry evident from the printout is the symmetry of the distribution about its peak: $h(r|100, 50, 10) = h(10 - r|100, 50, 10)$. However, this is not so in general; changing N to 99 results in a slightly unsymmetrical peak as we see from Table 3.3. The symmetric peak in Table 3.1 arises as follows: if we interchange M and $(N-M)$ and at the same time interchange r and $(n-r)$ we have in effect only interchanged the words "red" and "white", so the distribution is unchanged:

$$h(n-r|N, N-M, n) = h(r|N, M, n) \tag{3–26}$$

But when $M = N/2$, this reduces to the symmetry

$$h(n-r|N, M, n) = h(r|N, M, n)$$

observed in Table 3.1. By (3–25) the peak must be symmetric also when $n = N/2$.

r	$h(r)$	$H(r)$
0	0.000527	0.000527
1	0.006594	0.007121
2	0.035460	0.042581
3	0.108070	0.150651
4	0.206715	0.357367
5	0.259334	0.616700
6	0.216111	0.832812
7	0.118123	0.950934
8	0.040526	0.991461
9	0.007880	0.999341
10	0.000659	1.000000

Table 3.3: Hypergeometric Distribution, $N, M, n = 99, 50, 10$.

The hypergeometric distribution has two more symmetries not at all obvious intuitively or even visible in (3–18). Let us ask the robot for its probability $P(R_2|B)$ of red on the second draw. This is not the same calculation as (3–8), because the robot knows that, just prior to the second draw, there are only $(N-1)$ balls in the urn, not N. But it does not know what color of ball was removed on the first draw, so it does not know whether the number of red balls now in the urn is M or $(M-1)$. Then the basis for the Bernoulli urn result (3–5) is lost, and it might appear that the problem is indeterminate.

Yet it is quite determinate after all; the following is our first example of one of the useful techniques in probability calculations, which derives from the resolution of a proposition into disjunctions of simpler ones, as discussed in Chapters 1 and 2. The robot does know that either R_1 or W_1 is true, therefore a relation of Boolean algebra is

$$R_2 = (R_1 + W_1)R_2 = R_1R_2 + W_1R_2 \quad . \tag{3–27}$$

So we apply the sum rule and the product rule to get

$$\begin{aligned} P(R_2|B) &= P(R_1R_2|B) + P(W_1R_2|B) \\ &= P(R_2|R_1B)\,P(R_1|B) + P(R_2|W_1B)\,P(W_1|B)\ . \end{aligned} \tag{3–28}$$

But

$$P(R_2|R_1B) = \frac{M-1}{N-1}\ , \qquad P(R_2|W_1B) = \frac{M}{N-1} \tag{3–29}$$

and so

$$P(R_2|B) = \frac{M-1}{N-1}\,\frac{M}{N} + \frac{M}{N-1}\,\frac{N-M}{N} = \frac{M}{N}\,. \tag{3–30}$$

The complications cancel out, and we have the same probability of red on the first and second draws. Let us see whether this continues. For the third draw we have

$$R_3 = (R_1 + W_1)(R_2 + W_2)R_3 = R_1R_2R_3 + R_1W_2R_3 + W_1R_2R_3 + W_1W_2R_3 \tag{3–31}$$

and so

$$\begin{aligned} P(R_3|B) &= \frac{M}{N}\,\frac{M-1}{N-1}\,\frac{M-2}{N-2} + \frac{M}{N}\,\frac{N-M}{N-1}\,\frac{M-1}{N-2} \\ &+ \frac{N-M}{N}\,\frac{M}{N-1}\,\frac{M-1}{N-2} + \frac{N-M}{N}\,\frac{N-M-1}{N-1}\,\frac{M}{N-2} \\ &= \frac{M}{N}\,. \end{aligned} \tag{3–32}$$

Again all the complications cancel out. The robot's probability of red at any draw, *if it does not know the result of any other draw*, is always the same as the Bernoulli urn result (3–5). This is the first non–obvious symmetry. We shall not prove this in generality here, because it is contained as a special case of a more general result, Eq. (3–105) below.

The method of calculation illustrated by (3–28) and (3–31) is: resolve the quantity whose probability is wanted into mutually exclusive sub-propositions, then apply the sum rule and the product rule. If the sub-propositions are well chosen (*i.e.,* if they have some simple meaning in the context of the problem), their probabilities are often calculable. If they are not well chosen (as in the example of the penguins in the Comments at the end of Chapter 2), then of course this procedure cannot help us.

Logic Versus Propensity.

This suggests a new question. In finding the probability of red at the k'th draw, knowledge of what color was found at some earlier draw is clearly relevant because an earlier draw affects the number M_k of red balls in the urn for the k'th draw. Would knowledge of the color for a later draw be relevant? At first glance it seems that it could not be, because the result of a later draw cannot influence the value of M_k. For example, a well–known exposition of statistical mechanics (Penrose, 1979) takes it as a fundamental axiom that probabilities referring to the present time can depend only on what happened earlier, not on what happens later. The author considers this to be a necessary physical condition of "causality".

Therefore we stress again, as we did in Chapter 1, that inference is concerned with *logical* connections, which may or may not correspond to causal physical influences. To show why knowledge of later events is relevant to the probabilities of earlier ones, consider an urn which is known (background information B) to contain only one red and one white ball: $N = 2$, $M = 1$. Given only this information, the probability of red on the first draw is $P(R_1|B) = 1/2$. But then if the robot learns that red occurs on the second draw, it becomes certain that it did not occur on the first:

$$P(R_1|R_2B) = 0 \quad . \tag{3–33}$$

More generally, the product rule gives us

$$P(R_j R_k|B) = P(R_j|R_k B)\, P(R_k|B) = P(R_k|R_j B)\, P(R_j|B) \quad .$$

But we have just seen that $P(R_j|B) = P(R_k|B) = M/N$ for all j, k, so

$$P(R_j|R_k B) = P(R_k|R_j B) \quad , \qquad all\ j, k \quad . \tag{3–34}$$

Probability theory tells us that the results of later draws have precisely the same relevance as do the results of earlier ones! Even though performing the later draw does not physically affect the number M_k of red balls in the urn at the k'th draw, *information* about the result of a later draw has the same effect on our *state of knowledge* about what could have been taken on the k'th draw, as does information about an earlier one. This is our second non–obvious symmetry.

This result will be quite disconcerting to some schools of thought about the "meaning of probability". Although it is generally recognized that logical implication is not the same as physical causation, nevertheless there is a strong inclination to cling to the idea anyway, by trying to interpret a probability $P(A|B)$ as expressing some kind of partial causal influence of B on A. This is evident not only in the aforementioned work of Penrose, but more strikingly in the "propensity" theory of probability expounded by the philosopher Karl Popper (1957).

It appears to us that such a relation as (3–34) would be quite inexplicable from a propensity viewpoint, although the simple example (3–33) makes its logical necessity obvious. In any event, the theory of logical inference that we are developing here differs fundamentally, in outlook and in results, from the theory of physical causation envisaged by Penrose and Popper. It is evident that logical inference can be applied in many problems where assumptions of physical causation would not make sense.

This does not mean that we are forbidden to introduce the notion of "propensity" or physical causation; the point is rather that logical inference is applicable and useful whether or not a propensity exists. If such a notion (*i.e.,* that some such propensity exists) is formulated as a well–defined hypothesis, then our form of probability theory can analyze its implications. We shall do this in "Correction for Correlations" below. Also, we can test that hypothesis against alternatives in the light of the evidence, just as we can test any well–defined hypothesis. Indeed, one of the most common and important applications of probability theory is to decide whether there is evidence for a causal influence: is a new medicine more effective, or a new engineering design more reliable? Our study of hypothesis testing starts in Chapter 4.

In all the sciences, logical inference is more generally applicable. We agree that physical influences can propagate only forward in time; but logical inferences propagate equally well in either direction. An archaeologist uncovers an artifact that changes his knowledge of events thousands of years ago; were it otherwise, archaeology, geology, and paleontology would be impossible. The reasoning of Sherlock Holmes is also directed to inferring, from presently existing evidence, what events must have transpired in the past. The sounds reaching your ears from a marching band 600 meters distant change your state of knowledge about what the band was playing two seconds earlier.

As this suggests, and as we shall verify later, a fully adequate theory of nonequilibrium phenomena such as sound propagation, also requires that backward logical inferences be recognized and used, although they do not express physical causes. The point is that the best inferences we can make about any phenomenon – whether in physics, biology, economics, or any other field – must take into account all the relevant information we have, regardless of whether that information refers to times earlier or later than the phenomenon itself; this ought to be considered a platitude, not a paradox. At the end of this Chapter [Exercise (3.6)] the reader will have an opportunity to demonstrate this directly, by calculating a backward inference that takes into account a forward causal influence.

More generally, consider a probability distribution $p(x_1 \dots x_n|B)$, where x_i denotes the result of the i'th trial, and could take on, not just two values (red or white) but, say, the values $x_i = (1, 2, \dots, k)$ labelling k different colors. If the probability is invariant under any permutation of the x_i, then it depends only on the sample numbers $(n_1 \dots n_k)$ denoting how many times the result $x_i = 1$ occurs, how many times $x_i = 2$ occurs, *etc.* Such a distribution is called *exchangeable*; as we shall find later, exchangeable distributions have many interesting mathematical properties and important applications.

Returning to our Urn problem, it is clear already from the fact that the hypergeometric distribution is exchangeable, that every draw must have just the same relevance to every other draw regardless of their time order and regardless of whether they are near or far apart in the sequence. But this is not limited to the hypergeometric distribution; it is true of any exchangeable distribution (*i.e.*, whenever the probability of a sequence of events is independent of their order). So with a little more thought these symmetries, so inexplicable from the standpoint of physical causation, become obvious after all as propositions of logic.

Let us calculate this effect quantitatively. Supposing $j < k$, the proposition $R_j R_k$ (red at both draws j and k) is in Boolean algebra the same as

$$R_j R_k = (R_1 + W_1) \cdots (R_{j-1} + W_{j-1})\, R_j\, (R_{j+1} + W_{j+1}) \cdots (R_{k-1} + W_{k-1}) R_k \tag{3–35}$$

which we could expand in the manner of (3–31) into a logical sum of

$$2^{j-1} \cdot 2^{k-j-1} = 2^{k-2}$$

propositions, each specifying a full sequence, such as

$$W_1 R_2 W_3 \cdots R_j \cdots R_k \tag{3–36}$$

of k results. The probability $P(R_j R_k|B)$ is the sum of all their probabilities. But we know that, given B, the probability of any one sequence is independent of the order in which red and white appear. Therefore we can permute each sequence, moving R_j to the first position, and R_k to the second. That is, replace the sequence $(W_1 \cdots R_j \cdots)$ by $(R_1 \cdots W_j \cdots)$, *etc.* Recombining them, we have $(R_1 R_2)$ followed by every possible result for draws $(3, 4 \dots k)$. In other words, the probability of $R_j R_k$ is the same as that of

$$R_1 R_2 (R_3 + W_3) \cdots (R_k + W_k) = R_1 R_2 \tag{3–37}$$

and we have

$$P(R_j R_k|B) = P(R_1 R_2|B) = \frac{M(M-1)}{N(N-1)} \tag{3–38}$$

and likewise

$$P(W_j R_k|B) = P(W_1 R_2|B) = \frac{(N-M)M}{N(N-1)} \quad . \tag{3–39}$$

Therefore by the product rule

$$P(R_k|R_j B) = \frac{P(R_j R_k|B)}{P(R_j|B)} = \frac{M-1}{N-1} \tag{3–40}$$

and

$$P(R_k|W_j B) = \frac{P(W_j R_k|B)}{P(W_j|B)} = \frac{M}{N-1} \tag{3–41}$$

for all $j < k$. By (3–34), the results (3–40), (3–41) are true for all $j \neq k$.

Since as noted this conclusion appears astonishing to many people, we shall belabor the point by explaining it still another time in different words. The robot knows that the urn contained originally M red balls and $(N - M)$ white ones. Then learning that an earlier draw gave red, it knows that one less red ball is available for the later draws. The problem becomes the same as if we had started with an urn of $(N - 1)$ balls, of which $(M - 1)$ are red; (3–40) corresponds just to the solution (3–32) adapted to this different problem.

But why is knowing the result of a later draw equally cogent? Because if the robot knows that red will be drawn at any later time, then in effect one of the red balls in the urn must be "set aside" to make this possible. The number of red balls which could have been taken in earlier draws is reduced by one, as a result of having this information. The above example (3–33) is an extreme special case of this, where the conclusion is particularly obvious.

Reasoning from Less Precise Information

Now let us try to apply this understanding to a more complicated problem. Suppose the robot learns that red will be found at least once in later draws, but not at which draw or draws this will occur. That is, the new information is, as a proposition of Boolean algebra,

$$R_{later} \equiv R_{k+1} + R_{k+2} + \cdots + R_n \ . \tag{3–42}$$

This information reduces the number of red available for the k'th draw by at least one, but it is not obvious whether R_{later} has exactly the same implications as does R_n. To investigate this we appeal again to the symmetry of the product rule:

$$P(R_k R_{later}|B) = P(R_k|R_{later}B)\ P(R_{later}|B) = P(R_{later}|R_k B)\ P(R_k|B) \tag{3–43}$$

which gives us

$$P(R_k|R_{later}B) = P(R_k|B)\ \frac{P(R_{later}|R_k B)}{P(R_{later}|B)} \tag{3–44}$$

and all quantities on the right-hand side are easily calculated.

Seeing (3–42) one might be tempted to reason as follows:

$$P(R_{later}|B) = \sum_{j=k+1}^{n} P(R_j|B)$$

but this is not correct because, unless $M = 1$, the events R_j are not mutually exclusive, and as we see from (2–61), many more terms would be needed. This method of calculation would be very tedious.

To organize the calculation better, note that the denial of R_{later} is the statement that white occurs at all the later draws:

$$\overline{R}_{later} = W_{k+1}\ W_{k+2} \cdots W_n \ . \tag{3–45}$$

So $P(\overline{R}_{later}|B)$ is the probability of white at all the later draws, regardless of what happens at the earlier ones (*i.e.,* when the robot does not know what happens at the earlier ones). By exchangeability this is the same as the probability of white at the first $(n-k)$ draws, regardless of what happens at the later ones; from (3–12),

$$P(\overline{R}_{later}|B) = \frac{(N-M)!\,(N-n+k)!}{N!\,(N-M-n+k)!} = \binom{N-M}{n-k}\binom{N}{n-k}^{-1} . \qquad (3\text{–}46)$$

Likewise $P(\overline{R}_{later}|R_kB)$ is the same result for the case of $(N-1)$ balls, $(M-1)$ of which are red:

$$P(\overline{R}_{later}|R_kB) = \frac{(N-M)!}{(N-1)!}\,\frac{(N-n+k-1)!}{(N-M-n+k)!} = \binom{N-M}{n-k}\binom{N-1}{n-k}^{-1} \qquad (3\text{–}47)$$

Now (3–44) becomes

$$P(R_k|R_{later}B) = \frac{M}{N-n+k}\cdot\frac{\binom{N-1}{n-k}-\binom{N-M}{n-k}}{\binom{N}{n-k}-\binom{N-M}{n-k}} \qquad (3\text{–}48)$$

As a check, note that if $n = k+1$, this reduces to $(M-1)/(N-1)$, as it should.

At the moment, however, our interest in (3–48) is not so much in the numerical values, but in understanding the logic of the result. So let us specialize it to the simplest case that is not entirely trivial. Suppose we draw $n = 3$ times from an urn containing $N = 4$ balls, $M = 2$ of which are white, and ask how knowledge that red occurs at least once on the second and third draws, affects the probability of red at the first draw. This is given by (3–48) with $N = 4,\ M = 2,\ n = 3,\ k = 1$:

$$P(R_1|(R_2+R_3)B) = \frac{6-2}{12-2} = \frac{2}{5} = \frac{1}{2}\,\frac{1-\frac{1}{3}}{1-\frac{1}{6}} . \qquad (3\text{–}49)$$

The last form corresponding to (3–44). Compare this to the previously calculated probabilities:

$$P(R_1|B) = \frac{1}{2}\ , \qquad P(R_1|R_2B) = P(R_2|R_1B) = \frac{1}{3}\ .$$

What seems surprising is that

$$P(R_1|R_{later}B) > P(R_1|R_2B)\ . \qquad (3\text{–}50)$$

Most people guess at first that the inequality should go the other way; *i.e.,* knowing that red occurs at least once on the later draws ought to decrease the chances of red at the first draw more than does the information R_2. But in this case the numbers are so small that we can check the calculation (3–44) directly. To find $P(R_{later}|B)$ by the extended sum rule (2–61) now requires only one extra term:

$$\begin{aligned} P(R_{later}|B) &= P(R_2|B) + P(R_3|B) - P(R_2R_3|B) \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{2}\cdot\frac{1}{3} = \frac{5}{6}\ . \end{aligned} \qquad (3\text{–}51)$$

We could equally well resolve R_{later} into mutually exclusive propositions and calculate

$$\begin{aligned} P(R_{later}|B) &= P(R_2W_3|B) + P(W_2R_3|B) + P(R_2R_3|B) \\ &= \frac{1}{2}\cdot\frac{2}{3} + \frac{1}{2}\cdot\frac{2}{3} + \frac{1}{2}\cdot\frac{1}{3} = \frac{5}{6} \ . \end{aligned} \qquad (3\text{–}52)$$

The denominator $(1-1/6)$ in (3–49) has now been calculated in three different ways, with the same result. If the three results were not the same, we would have found an inconsistency in our rules, of the kind we sought to prevent by Cox's functional equation arguments in Chapter 2. This is a good example of what "consistency" means in practice, and it shows the trouble we would be in if our rules did not have it.

Likewise, we can check the numerator of (3–44) by an independent calculation:

$$\begin{aligned} P(R_{later}|R_1B) &= P(R_2|R_1B) + P(R_3|R_1B) - P(R_2R_3|R_1B) \\ &= \frac{1}{3} + \frac{1}{3} - \frac{1}{3}\cdot 0 = \frac{2}{3} \end{aligned} \qquad (3\text{–}53)$$

and the result (3–49) is confirmed. So we have no choice but to accept the inequality (3–50) and try to understand it intuitively. Let us reason as follows: The information R_2 reduces the number of red balls available for the first draw by one, and it reduces the number of balls in the urn available for the first draw by one, giving $P(R_1|R_2B) = (M-1)/(N-1) = \frac{1}{3}$. The information R_{later} reduces the "effective number of red balls" available for the first draw by more than one, but it reduces the number of balls in the urn available for the first draw by 2 (because it assures the robot that there are two later draws in which two balls are removed). So let us try tentatively to interpret the result (3–49) as

$$P(R_1|R_{later}B) = \frac{(M)_{eff}}{N-2} \qquad (3\text{–}54)$$

although we are not quite sure what this means. Given R_{later}, it is certain that at least one red ball is removed, and the probability that two are removed is by the product rule:

$$\begin{aligned} P(R_2R_3|R_{later}B) &= \frac{P(R_2R_3R_{later}|B)}{P(R_{later}|B)} = \frac{P(R_2R_3|B)}{P(R_{later}|B)} \\ &= \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{5}{6}} = \frac{1}{5} \end{aligned} \qquad (3\text{–}55)$$

because R_2R_3 implies R_{later}; *i.e.*, a relation of Boolean algebra is $(R_2R_3R_{later} = R_2R_3)$. Intuitively, given R_{later} there is probability $1/5$ that two red balls are removed, so the effective number removed is $1+(1/5) = 6/5$. The 'effective' number remaining for draw 1 is $4/5$. Indeed, (3–54) then becomes

$$P(R_1|R_{later}B) = \frac{4/5}{2} = \frac{2}{5} \qquad (3\text{–}56)$$

in agreement with our better motivated but less intuitive calculation (3–49).

Expectations.

Another way of looking at this result appeals more strongly to our intuition and generalizes far beyond the present problem. We can hardly suppose that the reader is not already familiar with the idea of expectation, but this is the first time it has appeared in the present work, so we pause to define it. If a variable quantity X can take on the particular values $(x_1, x_2 \cdots x_n)$ in n mutually exclusive and exhaustive situations and the robot assigns corresponding probabilities $(p_1, p_2 \cdots p_n)$ to them, then the quantity

$$\langle X \rangle = E(X) = \sum_{i=1}^{n} p_i x_i \tag{3–57}$$

is called the *expectation* (in the older literature, *mathematical expectation* or *expectation value*) of X. It is a weighted average of the possible values, weighted according to their probabilities. Statisticians and mathematicians generally use the notation $E(X)$; but physicists, having already pre–empted E to stand for energy and electric field, use the bracket notation $\langle X \rangle$. We shall use both notations here; they have the same meaning but sometimes one is easier to read than the other.

Like most of the standard terms that arose out of the distant past, the term "expectation" seems singularly inappropriate to us; for it is almost never a value that anyone "expects" to find. Indeed, it is often known to be an impossible value. But we adhere to it because of centuries of precedent.

Given R_{later}, what is the expectation of the number of red balls in the urn for draw #1? There are three mutually exclusive possibilities compatible with R_{later}:

$$R_2 W_3, \; W_2 R_3, \; R_2 R_3$$

for which M is $(1, 1, 0)$ respectively, and for which the probabilities are as in (3–55), (3–56):

$$\begin{aligned} P(R_2 W_3 | R_{later} B) &= \frac{P(R_2 W_3 | B)}{P(R_{later}|B)} = \frac{(1/2)\cdot(2/3)}{(5/6)} = \frac{2}{5} \\ P(W_2 R_3 | R_{later} B) &= \frac{2}{5} \\ P(R_2 R_3 | R_{later} B) &= \frac{1}{5} \end{aligned} \tag{3–58}$$

So

$$\langle M \rangle = 1 \cdot \frac{2}{5} + 1 \cdot \frac{2}{5} + 0 \cdot \frac{1}{5} = \frac{4}{5} \tag{3–59}$$

Thus what we called intuitively the "effective" value of M in (3–54) is really the expectation of M.

We can now state (3–54) in a more cogent way: when the fraction $F = M/N$ of red balls is known, then the Bernoulli urn rule applies and $P(R_1|B) = F$. When F is unknown, the probability of red is the expectation of F:

$$P(R_1|B) = \langle F \rangle = E(F) \ . \tag{3–60}$$

If M and N are both unknown, the expectation is over the joint probability distribution for M and N.

That a probability is numerically equal to the expectation of a fraction will prove to be a general rule that holds as well in thousands of far more complicated situations, providing one of the most useful and common rules for physical prediction. We leave it as an exercise for the reader to show that the more general result (3–48) can also be calculated in the way suggested by (3–60).

Other Forms and Extensions.

The hypergeometric distribution (3–18) can be written in various ways. The nine factorials can be organized into binomial coefficients also as follows:

$$h(r; N, M, n) = \frac{\binom{n}{r}\binom{N-n}{M-r}}{\binom{N}{M}} \tag{3–61}$$

But the symmetry under exchange of M and n is still not evident; to see it one must write out (3–18) or (3–61) in full, displaying all the individual factorials.

We may also rewrite (3–18), as an aid to memory, in a more symmetric form: the probability of drawing exactly r red balls and w white ones in $n = r + w$ draws from an urn containing R red and W white, is

$$h(r) = \frac{\binom{R}{r}\binom{W}{w}}{\binom{R+W}{r+w}} \tag{3–62}$$

and in this form it is easily generalized. Suppose that instead of only two colors, there are k different colors of balls, in the urn, N_1 of color 1, N_2 of color 2,. . ., N_k of color k. The probability of drawing r_1 balls of color 1, r_2 of color 2,. . ., r_k of color k in $n = \Sigma r_i$ draws is, as the reader may verify, the generalized hypergeometric distribution:

$$h(r_1 \cdots r_k | N_1 \cdots N_k) = \frac{\binom{N_1}{r_1} \cdots \binom{N_k}{r_k}}{\binom{\Sigma N_i}{\Sigma r_i}} \tag{3–63}$$

Probability as a Mathematical Tool.

From the result (3–63) one may obtain a number of identities obeyed by the binomial coefficients. For example, we may decide not to distinguish between colors 1 and 2; *i.e.,* a ball of either color is declared to have color "a". Then from (3–63) we must have on the one hand,

$$h(r_a, r_3 \cdots r_k | N_a N_3 \cdots N_k) = \frac{\binom{N_a}{r_a}\binom{N_3}{r_3} \cdots \binom{N_k}{r_k}}{\binom{\Sigma N_i}{\Sigma r_i}} \tag{3–64}$$

with

$$N_a = N_1 + N_2 \ , \qquad\qquad r_a = r_1 + r_2 \ . \tag{3–65}$$

But the event r_a can occur for any values of r_1, r_2 satisfying (3–65), and so we must have also, on the other hand,

$$h(r_a, r_3 \cdots r_k | N_a N_3 \cdots N_k) = \sum_{r_1=0}^{r_a} h(r_1, r_a - r_1, r_3 \cdots r_k | N_1 \cdots N_k) \ . \tag{3–66}$$

Then, comparing (3–64) and (3–66) we have the identity

$$\binom{N_a}{r_a} = \sum_{r_1=0}^{r_a} \binom{N_1}{r_1} \binom{N_2}{r_a - r_1} . \tag{3–67}$$

Continuing in this way, we can derive a multitude of more complicated identities obeyed by the binomial coefficients. For example,

$$\binom{N_1 + N_2 + N_3}{r_a} = \sum_{r_1=0}^{r_a} \sum_{r_2=0}^{r_1} \binom{N_1}{r_1} \binom{N_2}{r_2} \binom{N_3}{r_a - r_1 - r_2} . \tag{3–68}$$

In many cases, probabilistic reasoning is a powerful tool for deriving purely mathematical results; more examples of this are given by Feller (1951, Chapters 2, 3) and in later Chapters of the present work.

The Binomial Distribution.

Although somewhat complicated mathematically, the hypergeometric distribution arises from a problem that is very clear and simple conceptually; there are only a finite number of possibilities and all the above results are exact for the problems as stated. As an introduction to a mathematically simpler, but conceptually far more difficult problem, we examine a limiting form of the hypergeometric distribution.

The complication of the hypergeometric distribution arises because it is taking into account the changing contents of the urn; knowing the result of any draw changes the probability of red for any other draw. But if the number N of balls in the urn is very large compared to the number drawn ($N >> n$), then this probability changes very little, and in the limit $N \to \infty$ we should have a simpler result, free of such dependences. To verify this, we write the hypergeometric distribution (3–18) as

$$h(r; N, M, n) = \frac{\left[\frac{1}{N^r}\binom{M}{r}\right] \left[\frac{1}{N^{n-r}}\binom{N-M}{n-r}\right]}{\left[\frac{1}{N^n}\binom{N}{n}\right]} . \tag{3–69}$$

The first factor is

$$\frac{1}{N^r}\binom{M}{r} = \frac{1}{r!}\,\frac{M}{N}\left(\frac{M}{N} - \frac{1}{N}\right)\left(\frac{M}{N} - \frac{2}{N}\right)\cdots\left(\frac{M}{N} - \frac{r-1}{N}\right) \tag{3–70}$$

and in the limit $N \to \infty,\ M \to \infty,\ M/N \to f$ we have

$$\frac{1}{N^r}\binom{M}{r} \to \frac{f^r}{r!} \tag{3–71}$$

Likewise

$$\frac{1}{N^{n-r}}\binom{M-1}{n-r} \to \frac{(1-f)^{n-r}}{(n-r)!} \tag{3–72}$$

$$\frac{1}{N^n}\binom{N}{n} \to \frac{1}{n!} . \tag{3–73}$$

In principle we should, of course, take the limit of the product in (3–69), not the product of the limits. But in (3–69) we have defined the factors so that each has its own independent limit, so the result is the same; the hypergeometric distribution goes into

$$h(r; N, M, n) \to b(r|n, f) \equiv \binom{n}{r} f^r (1-f)^{n-r} \tag{3–74}$$

called the *binomial* distribution, because evaluation of the generating function (3–20) now reduces to

$$G(t) \equiv \sum_{r=0}^{n} b(r|n, f)\, t^r = (1 - f + ft)^n \ , \tag{3–75}$$

an example of Newton's binomial theorem. The program BINOMIAL.BAS in Appendix I calculates $b(r|n, f)$ for most values of n, f likely to be of interest in applications.

Fig. 3.1 compares three hypergeometric distributions calculated by HYPERGEO.BAS with $N = 15, 30, 100$ and $M/N = 0.4$, $n = 10$ to the binomial distribution with $n = 10$, $f = 0.4$ calculated by BINOMIAL.BAS. All have their peak at $r = 4$, and all distributions have the same first moment $\langle r \rangle = E(r) = 4$, but the binomial distribution is broader.

space for Fig. 3.1

The $N = 15$ hypergeometric distribution is zero for $r = 0$ and $r > 6$, since on drawing 10 balls from an urn containing only 6 red and 9 white, it is not possible to get fewer than one or more than 6 red balls. When $N > 100$ the hypergeometric distribution agrees so closely with the binomial that for most purposes it would not matter which one we used. Analytical properties of the binomial distribution are collected in Appendix E.

We can carry out a similar limiting process on the generalized hypergeometric distribution (3–63). It is left as an exercise to show that in the limit where all $N_i \to \infty$ in such a way that the fractions

$$f_i \equiv \frac{N_i}{\Sigma N_i} \tag{3–76}$$

tend to constants, (3–63) goes into the *multinomial distribution*

$$m(r_1 \cdots r_k | f_1 \cdots f_k) = \frac{r!}{r_1! \cdots r_k!} \, f_1^{r_1} \cdots f_k^{r_k} \ , \tag{3–77}$$

where $r \equiv \Sigma r_i$. And, as in (3–75) we can define a generating function of $(k-1)$ variables, from which we can prove that (3–77) is correctly normalized, and derive many other useful results.

Exercise 3.2. Probability of a Full Set. Suppose an urn contains $N = \sum N_i$ balls, N_1 of color 1, N_2 of color 2, $\cdots$ N_k of color k. We draw m balls without replacement; what is the probability that we have at least one of each color? Supposing $k = 5$, all $N_i = 10$, how many do we need to draw in order to have at least a 90% probability of getting a full set?

Exercise 3.3. Reasoning Backwards. Suppose that in the previous exercise k is initially unknown, but we know that the urn contains exactly 50 balls. Drawing out 20 of them, we find 3 different colors; now what do we know about k? We know from deductive reasoning (*i.e.*, with certainty) that $3 \le k \le 33$; but can you set narrower limits $k_1 \le k \le k_2$ within which it is highly likely to be? [*Hint: this question goes beyond the sampling theory of this Chapter because, like most real scientific problems, the answer depends to some degree on our common sense judgments; nevertheless our rules of probability theory are quite capable of dealing with it, and persons with reasonable common sense cannot differ appreciably in their conclusions*].

Exercise 3.4. Matching. The M urns are now numbered 1 to M, and M balls, also numbered 1 to M, are thrown into them, one in each urn. If the numbers of a ball and its urn are the same, we have a match. Show that the probability of at least one match is

$$P = \sum_{k=1}^{M} (-1)^{k+1}/k!$$

As $M \to \infty$, this converges to $1 - 1/e = 0.632$. The result is surprising to many, because however large M is, there remains an appreciable probability of no match at all.

Exercise 3.5. Occupancy. N balls are tossed into M urns; there are evidently M^N ways this can be done. If the robot considers them all equally likely, what is its probability that each urn receives at least one ball?

Sampling With Replacement

Up to now, we have considered only the case where we sample without replacement; and that is evidently appropriate for many real situations. For example, in a quality control application, what we have called simply "drawing a ball" might consist really of taking a manufactured item such as an electric light bulb from a carton of them and testing it to destruction. In a chemistry experiment it might consist of weighing out a sample of an unknown protein, then dissolving it in hot sulfuric acid to measure its nitrogen content. In either case, there can be no thought of "drawing that same ball" again.

But suppose now that, being less destructive, we sample balls from the urn and, after recording the "color" (*i.e.*, the relevant property) of each, we replace it in the urn before drawing the next ball.

This case, of sampling with replacement, is enormously more complicated conceptually, but with some assumptions usually made, ends up being simpler mathematically, than sampling without replacement. For, let us go back to the probability of drawing two red balls in succession. Denoting by B' the same background information as before except for the added stipulation that the balls are to be replaced, we still have an equation like (3–9):

$$P(R_1 R_2|B') = P(R_1|B')\, P(R_2|R_1 B') \tag{3–78}$$

and the first factor is still, evidently, (M/N); but what is the second one?

Answering this would be, in general, a very difficult problem requiring much additional analysis if the background information B' includes some simple but highly relevant common–sense information that we all have. What happens to that red ball that we put back in the urn? If we merely dropped it into the urn, and immediately drew another ball, then it was left lying on the top of the other balls, (or in the top layer of balls); and so it is more likely to be drawn again than any other specified ball, whose location in the urn is unknown. But this upsets the whole basis of our calculation, because the probability of drawing any particular (i'th) ball is no longer given by the Bernoulli Urn Rule which led to (3–10).

Digression: A Sermon on Reality vs. Models

The difficulty we face here is that many things which were irrelevant from symmetry as long as the robot's state of knowledge was invariant under any permutation of the balls, suddenly become relevant, and by one of our desiderata of rationality, the robot must take into account all the relevant information it has. But the probability of drawing any particular ball now depends on such details as the exact size and shape of the urn, the size of the balls, the exact way in which the first one was tossed back in, the elastic properties of balls and urn, the coefficients of friction between balls and between ball and urn, the exact way you reach in to draw the second ball, *etc.* In a symmetric situation, all of these details are irrelevant.

But even if all these relevant data were at hand, we do not think that a team of the world's best scientists and mathematicians, backed up by all the world's computing facilities, would be able to solve the problem; or would even know how to get started on it. Still, it would not be quite right to say that the problem is unsolvable *in principle*; only so complicated that it is not worth anybody's time to think about it. So what do we do?

In probability theory there is a very clever trick for handling a problem that becomes too difficult. We just solve it anyway by:

(1) Making it still harder;
(2) Redefining what we mean by "solving" it, so that it becomes something we *can* do;
(3) Inventing a dignified and technical–sounding word to describe this procedure, which has the psychological effect of concealing the real nature of what we have done, and making it appear respectable.

In the case of sampling with replacement, we apply this strategy by

(1) Supposing that after tossing the ball in, we shake up the urn. However complicated the problem was initially, it now becomes many orders of magnitude more complicated, because the solution now depends on every detail of the precise way we shake it, in addition to all the factors mentioned above;
(2) Asserting that the shaking has somehow made all these details irrelevant, so that the problem reverts back to the simple one where the Bernoulli Urn Rule applies;
(3) Inventing the dignified–sounding word *randomization* to describe what we have done. This term is, evidently, a euphemism whose real meaning is: *deliberately throwing away relevant information when it becomes too complicated for us to handle.*

We have described this procedure in laconic terms, because an antidote is needed for the impression created by some writers on probability theory, who attach a kind of mystical significance to it. For some, declaring a problem to be "randomized" is an incantation with the same purpose and effect as those uttered by an exorcist to drive out evil spirits; *i.e.,* it cleanses their subsequent calculations and renders them immune to criticism. We agnostics often envy the True Believer, who thus acquires so easily that sense of security which is forever denied to us.

However, in defense of this procedure, we have to admit that it often leads to a useful approximation to the correct solution; *i.e.,* the complicated details, while undeniably relevant in principle, might nevertheless have little numerical effect on the answers to certain particularly simple questions, such as the probability of drawing r red balls in n trials when n is sufficiently small. But from the standpoint of principle, an element of vagueness necessarily enters at this point; for while we may feel intuitively that this leads to a good approximation, we have no proof of this, much less a reliable estimate of the accuracy of the approximation, which presumably improves with more shaking.

The vagueness is evident particularly in the fact that different people have widely divergent views about how much shaking is required to justify step (2). Witness the minor furor surrounding a Government–sponsored and nationally televised game of chance some years ago, when someone objected that the procedure for drawing numbers from a fish bowl to determine the order of call–up of young men for Military Service was "unfair" because the bowl hadn't been shaken enough to make the drawing "truly random," whatever that means. Yet if anyone had asked the objector: "To *whom* is it unfair?" he could not have given any answer except, "To those whose numbers are on top; I don't know who they are." But after any amount of further shaking, this will still be true! So what does the shaking accomplish?

Shaking does not make the result "random", because that term is basically meaningless as an attribute of the real world; it has no clear definition applicable in the real world. The belief that "randomness" is some kind of real property existing in Nature is a form of the Mind Projection Fallacy which says, in effect, "I don't know the detailed causes – *therefore* – Nature does not know them." What shaking accomplishes is very different. It does not affect *Nature's* workings in any way; it only ensures that no *human* is able to exert any wilful influence on the result. Therefore nobody can be charged with "fixing" the outcome.

At this point, you may accuse us of nit–picking, because you know that after all this sermonizing, we are just going to go ahead and use the randomized solution like everybody else does. Note, however, that our objection is not to the procedure itself, provided that we acknowledge honestly what we are doing; *i.e.,* instead of solving the real problem, we are making a practical compromise and being, of necessity, content with an approximate solution. That is something we have to do in all areas of applied mathematics, and there is no reason to expect probability theory to be any different.

Our objection is to this belief that by randomization we somehow make our subsequent equations exact; so exact that we can then subject our solution to all kinds of extreme conditions and believe the results, applied to the real world. The most serious and most common error resulting from this belief is in the derivation of limit theorems (*i.e.,* when sampling with replacement, nothing prevents us from passing to the limit $n \to \infty$ and obtaining the usual "laws of large numbers"). If we do not recognize the approximate nature of our starting equations, we delude ourselves into believing that we have proved things (such as the identity of probability and limiting frequency) that are just not true in real repetitive experiments.

The danger here is particularly great because mathematicians generally regard these limit theorems as the most important and sophisticated fruits of probability theory, and have a tendency to use language which implies that they are proving properties of the real world. Our point is that these theorems are valid properties *of the abstract mathematical model that was defined and analyzed*. The issue is: to what extent does that model resemble the real world? It is probably safe to say that no limit theorem is directly applicable in the real world, simply because no mathematical model captures every circumstance that is relevant in the real world. The person who believes that he is proving things about the real world, is a victim of the Mind Projection Fallacy.

Back to the Problem. Returning to the equations, what answer can we now give to the question posed after Eq. (3–78)? The probability $P(R_2|R_1B')$ of drawing a red ball on the second draw, clearly depends not only on N and M, but also on the fact that a red one has already been drawn and replaced. But this latter dependence is so complicated that we can't, in real life, take it into account; so we shake the urn to "randomize" the problem, and then declare R_1 to be irrelevant: $P(R_2|R_1B') = P(R_2|B') = M/N$. After drawing and replacing the second ball, we again shake the urn, declare it "randomized", and set $P(R_3|R_2R_1B') = P(R_3|B') = M/N$, etc. In this approximation, the probability of drawing a red one at *any* trial, is (M/N).

But this is not just a repetition of what we learned in (3–32); what is new here is that the result now holds *whatever information the robot may have about what happened in the other trials.* This leads us to write the probability of drawing exactly r red balls in n trials regardless of order, as

$$\binom{n}{r} \left(\frac{M}{N}\right)^r \left(\frac{N-M}{N}\right)^{n-r} \qquad (3\text{–}79)$$

which is just the binomial distribution (3–74). Randomized sampling with replacement from an urn with finite N has approximately the same effect as passage to the limit $N \to \infty$ without replacement.

Evidently, for small n, this approximation will be quite good; but for large n these small errors can accumulate (depending on exactly how we shake the urn, etc.) to the point where (3–79) is misleading. Let us demonstrate this by a simple, but realistic, extension of the problem.

Correction for Correlations

Suppose that, from an intricate logical analysis, drawing and replacing a red ball increases the probability of a red one at the next draw by some small amount $\epsilon > 0$, while drawing and replacing a white one decreases the probability of a red one at the next draw by a (possibly equal) small quantity $\delta > 0$; and that the influence of earlier draws than the last one is negligible compared to ϵ or δ. You may call this effect a small "propensity" if you like; at least it expresses a physical causation that operates only forward in time. Then, letting C stand for all the above background information including the statements just made about correlations, and the information that we draw n balls, we have

$$P(R_k|R_{k-1}, C) = p + \epsilon\ , \qquad P(R_k|W_{k-1}, C) = p - \delta$$

$$(3\text{–}80)$$

$$P(W_k|R_{k-1}, C) = 1 - p - \epsilon\ , \qquad P(W_k|W_{k-1}, C) = 1 - p + \delta$$

where $p \equiv M/N$. From this, the probability of drawing r red, $(n-r)$ white balls in any specified order, is easily seen to be:

$$p(p+\epsilon)^c (p-\delta)^{c'} (1-p+\delta)^w (1-p-\epsilon)^{w'} \qquad (3\text{–}81)$$

if the first draw is red, while if the first is white, the first factor in (3–81) should be $(1-p)$. Here c is the number of red draws preceded by red ones, c' the number of red preceded by white, w the number of white draws preceded by white, and w' the number of white preceded by red. Evidently,

$$c + c' = \left\{ \begin{matrix} r-1 \\ r \end{matrix} \right\} , \qquad w + w' = \left\{ \begin{matrix} n-r \\ n-r-1 \end{matrix} \right\} \tag{3–82}$$

the upper case and lower cases holding when the first draw is red or white, respectively.

When r and $(n-r)$ are small, the presence of ϵ and δ in (3–81) makes little difference, and it reduces for all practical purposes to

$$p^r (1-p)^{n-r} \tag{3–83}$$

as in the binomial distribution (3–79). But as these numbers increase, we can use relations of the form

$$\left(1 + \frac{\epsilon}{p}\right)^c \simeq \exp\left(\frac{\epsilon c}{p}\right) \tag{3–84}$$

and (3–81) goes into

$$p^r (1-p)^{n-r} \exp\left\{ \frac{\epsilon c - \delta c'}{p} + \frac{\delta w - \epsilon w'}{1-p} \right\} \tag{3–85}$$

The probability of drawing r red, $(n-r)$ white balls now depends on the order in which red and white appear, and for a given ϵ, when the numbers c, c', w, w' become sufficiently large, the probability can become arbitrarily large (or small) compared to (3–79).

We see this effect most clearly if we suppose that $N = 2M$, $p = 1/2$, in which case we will surely have $\epsilon = \delta$. The exponential factor in (3–85) then reduces to:

$$\exp\{2\epsilon[(c-c') + (w-w')]\} \tag{3–86}$$

This shows that (1) as the number n of draws tends to infinity, the probability of results containing "long runs"; *i.e.,* long strings of red (or white) balls in succession, becomes arbitrarily large compared to the value given by the "randomized" approximation; (2) this effect becomes appreciable when the numbers (ϵc), etc., become of order unity. Thus, if $\epsilon = 10^{-2}$, the randomized approximation can be trusted reasonably well as long as $n < 100$; beyond that, we might delude ourselves by using it. Indeed, it is notorious that in real repetitive experiments where conditions appear to be the same at each trial, such runs – although extremely improbable on the randomized approximation – are nevertheless observed to happen.

Now let us note how the correlations expressed by (3–80) affect some of our previous calculations. The probabilities for the first draw are of course the same as (3–8); now use the notation

$$p = P(R_1|C) = \frac{M}{N} , \qquad q = 1 - p = P(W_1|C) = \frac{N-M}{N} . \tag{3–87}$$

But for the second trial we have instead of (3–30)

$$\begin{aligned} P(R_2|C) &= P(R_2 R_1|C) + P(R_2 W_1|C) \\ &= P(R_2|R_1 C)\, P(R_1|C) + P(R_2|W_1 C)\, P(W_1|C) \\ &= (p+\epsilon)p + (p-\delta)q \\ &= p + (p\epsilon - q\delta) \end{aligned} \tag{3–88}$$

and continuing for the third trial,

$$\begin{aligned} P(R_3|C) &= P(R_3|R_2 C)\, P(R_2|C) + P(R_3|W_2 C)\, P(W_2|C) \\ &= (p+\epsilon)(p + p\epsilon - q\delta) + (p-\delta)(q - p\epsilon + q\delta) \\ &= p + (1+\epsilon+\delta)(p\epsilon - q\delta)\ . \end{aligned} \tag{3–89}$$

We see that $P(R_k|C)$ is no longer independent of k; the correlated probability distribution is no longer exchangeable. But does $P(R_k|C)$ approach some limit as $k \to \infty$?

It would be almost impossible to guess the general $P(R_k|C)$ by induction, following the method (3–88), (3–89) a few steps further. For this calculation we need a more powerful method. If we write the probabilities for the k'th trial as a vector

$$V_k \equiv \begin{pmatrix} P(R_k|C) \\ P(W_k|C) \end{pmatrix} \tag{3–90}$$

then Equation (3–80) can be expressed in matrix form:

$$V_k = M V_{k-1}\ , \tag{3–91}$$

with

$$M = \begin{pmatrix} (p+\epsilon) & (p-\delta) \\ (q-\epsilon) & (q+\delta) \end{pmatrix}\ . \tag{3–92}$$

This defines a *Markov chain* of probabilities, and M is called the *transition matrix.* Now the slow induction of (3–88), (3–89) proceeds instantly to any distance we please:

$$V_k = M^{k-1} V_1\ . \tag{3–93}$$

So to have the general solution, we need only to find the eigenvectors and eigenvalues of M. The characteristic polynomial is

$$C(\lambda) \equiv \det(M_{ij} - \lambda\delta_{ij}) = \lambda^2 - \lambda(1+\epsilon+\delta) + (\epsilon+\delta) \tag{3–94}$$

so the roots of $C(\lambda) = 0$ are the eigenvalues

$$\begin{aligned} \lambda_1 &= 1 \\ \lambda_2 &= \epsilon + \delta\ . \end{aligned} \tag{3–95}$$

Now for any 2×2 matrix

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{3–96}$$

with an eigenvalue λ, the corresponding (non–normalized) right eigenvector is

$$x = \begin{pmatrix} b \\ \lambda - a \end{pmatrix} \tag{3–97}$$

for which we have at once $Mx = \lambda x$. Therefore, our eigenvectors are

$$x_1 = \begin{pmatrix} p-\delta \\ q-\epsilon \end{pmatrix}, \qquad x_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} . \tag{3–98}$$

These are not orthogonal, since M is not a symmetric matrix. Nevertheless, if we use (3–98) to define the transformation matrix

$$S = \begin{pmatrix} (p-\delta) & 1 \\ (q-\epsilon) & -1 \end{pmatrix} \tag{3–99}$$

we find its inverse to be

$$S^{-1} = \frac{1}{1-\epsilon-\delta} \begin{pmatrix} 1 & 1 \\ (q-\epsilon) & -(p-\delta) \end{pmatrix} \tag{3–100}$$

and we can verify by direct matrix multiplication that

$$S^{-1}MS = \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \tag{3–101}$$

where Λ is the diagonalized matrix. Then we have for any r, positive, negative, or even complex:

$$M^r = S\Lambda^r S^{-1} \tag{3–102}$$

or,

$$M^r = \frac{1}{1-\epsilon-\delta} \begin{pmatrix} (p-\delta)+(\epsilon+\delta)^r(q-\epsilon) & (p-\delta)[1-(\epsilon+\delta)^r] \\ (q-\epsilon)[1-(\epsilon+\delta)^r] & (q-\epsilon)+(\epsilon+\delta)^r(p-\delta) \end{pmatrix} \tag{3–103}$$

and since

$$V_1 = \begin{pmatrix} p \\ q \end{pmatrix} \tag{3–104}$$

the general solution (3–93) sought is

$$P(R_k|C) = \frac{(p-\delta)-(\epsilon+\delta)^{k-1}(p\epsilon-q\delta)}{1-\epsilon-\delta} . \tag{3–105}$$

We can check that this agrees with (3–87), (3–88), (3–89). From examining (3–105) it is clear why it would have been almost impossible to guess the general formula by induction. When $\epsilon = \delta = 0$, this reduces to $P(R_k|C) = p$, supplying the proof promised after Eq. (3–32).

Although we started this discussion by supposing that ϵ and δ were small and positive, we have not actually used that assumption and so, whatever their values, the solution (3–105) is exact for the abstract model that we have defined. This enables us to include two interesting extreme cases. If not small, ϵ and δ must be at least bounded, because all quantities in (3–80) must be probabilities (that is, in $[0, 1]$). This requires that

$$-p \le \epsilon \le q \ , \qquad -q \le \delta \le p \tag{3–106}$$

or

$$-1 \le \epsilon + \delta \le 1 \ . \tag{3–107}$$

But from (3–106), $\epsilon + \delta = 1$ if and only if $\epsilon = q$, $\delta = p$, in which case the transition matrix reduces to the unit matrix

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{3–108}$$

and there are no "transitions". This is a degenerate case in which the positive correlations are so strong that whatever color happens to be drawn on the first trial, is certain to be drawn also on all succeeding ones:

$$P(R_k|C) = p \ , \qquad all\ k \ . \tag{3–109}$$

Likewise, if $\epsilon + \delta = -1$, then the transition matrix must be

$$M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{3–110}$$

and we have nothing but transitions; *i.e.,* the negative correlations are so strong that the colors are certain to alternate after the first draw:

$$P(R_k|C) = \begin{Bmatrix} p, & k \text{ odd} \\ q, & k \text{ even} \end{Bmatrix} \ . \tag{3–111}$$

This case is unrealistic because intuition tells us rather strongly that ϵ and δ should be positive quantities; surely, whatever the logical analysis used to assign the numerical value of ϵ, leaving a red ball in the top layer must *increase*, not decrease, the probability of red on the next draw. But if ϵ and δ must not be negative, then the lower bound in (3–107) is really zero, which is achieved only when $\epsilon = \delta = 0$. Then M in (3–92) becomes singular, and we revert to the binomial distribution case already discussed.

In the intermediate and realistic cases where $0 < |\epsilon + \delta| < 1$, the last term of (3–105) attenuates exponentially with k, and in the limit

$$P(R_k|C) \to \frac{p - \delta}{1 - \epsilon - \delta} \ . \tag{3–112}$$

But although these single–trial probabilities settle down to steady values as in an exchangeable distribution, the underlying correlations are still at work and the limiting distribution is not exchangeable. To see this, let us consider the conditional probabilities $P(R_k|R_jC)$. These are found by noting that the Markov chain relation (3–91) holds whatever the vector V_{k-1}; *i.e.,* whether or not it is the vector generated from V_1 as in (3–93). Therefore, if we are given that red occurred on the j'th trial, then

$$V_j = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

and we have from (3–91)

$$V_k = M^{k-j}\, V_j \ , \qquad\qquad j \le k \tag{3–113}$$

from which, using (3–102),

$$P(R_k|R_jC) = \frac{(p-\delta) + (\epsilon+\delta)^{k-j}\,(q-\epsilon)}{1-\epsilon-\delta} \quad , \qquad j < k \ . \tag{3–114}$$

which approaches the same limit (3–112). The forward inferences are about what we might expect; the steady value (3–112) plus a term that decays exponentially with distance. But the backward inferences are different; note that the general product rule holds, as always:

$$P(R_kR_j|C) = P(R_k|R_jC)\,P(R_j|C) = P(R_j|R_kC)\,P(R_k|C) \,. \tag{3–115}$$

Therefore, since we have seen that $P(R_k|C) \neq P(R_j|C)$, it follows that

$$P(R_j|R_kC) \neq P(R_k|R_jC) \ . \tag{3–116}$$

The backward inference is still possible, but it is no longer the same formula as the forward inference as it would be in an exchangeable sequence.

As we shall see later, this example is the simplest possible 'baby' version of a very common and important physical problem; an irreversible process in the 'Markovian approximation'. Another common technical language would call it an *autoregressive model* of first order. It can be generalized greatly to the case of matrices of arbitrary dimension and many–step or continuous, rather than single–step, memory influences. But for reasons noted earlier (confusion of inference and causality in the literature of statistical mechanics) the backward inference part of the solution is almost always missed. Some try to do backward inference by extrapolating the forward solution backward in time, with quite bizarre and unphysical results. Therefore the reader is, in effect, conducting new research in doing the following exercise.

Exercise (3.6) Find the explicit formula $P(R_j|R_kC)$ for the backward inference corresponding to the result (3–114) by using (3–105) and (3–115). Then (a) Explain the reason for the difference between forward and backward inferences in simple intuitive terms. (b) In what way does the backward inference differ from the forward inference extrapolated backward? Which is more reasonable intuitively? (c) Do backward inferences also decay to steady values? If so, is a property somewhat like exchangeability restored for events sufficiently separated? For example, if we consider only every tenth draw or every hundredth draw, do we approach an exchangeable distribution on this subset?

Simplification

The above formulas (3–87) – (3–118) hold for any ϵ, δ satisfying the inequalities (3–106). But surveying them, we note that a remarkable simplification occurs if they satisfy

$$p\epsilon = q\delta \ . \tag{3–117}$$

For then we have

$$\frac{p-\epsilon}{1-\epsilon-\delta} = p \ , \qquad \frac{q-\delta}{1-\epsilon-\delta} = q \ , \qquad \epsilon+\delta = \frac{\epsilon}{q} \tag{3–118}$$

and our main results (3–105), (3–114) collapse to

$$P(R_k|C) = p \quad , \quad all\ k \tag{3–119}$$

$$P(R_k|R_jC) = P(R_j|R_kC) = p + q\left(\frac{\epsilon}{q}\right)^{|k-j|} \quad , \qquad all\ k, j \ . \tag{3–120}$$

The distribution is still not exchangeable, since the conditional probabilities (3–120) still depend on the separation $|k-j|$ of the trials; but the symmetry of forward and backward inferences is restored even though the causal influences ϵ, δ operate only forward. Indeed, we see from our derivation of (3–34) that this forward – backward symmetry is a necessary consequence of (3–119) whether or not the distribution is exchangeable.

What is the meaning of this magic condition (3–117)? It does not make the matrix M assume any particularly simple form, and it does not turn off the effect of the correlations. What it does is to make the solution (3–119) invariant; that is, the initial vector (3–104) is then equal but for normalization to the eigenvector x_1 in (3–98), so the initial vector remains unchanged by the matrix (3–92).

In general, of course, there is no reason why this simplifying condition should hold. Yet in the case of our urn, we can see a kind of rationale for it. Suppose that when the urn has initially N balls, they are in L layers. Then after withdrawing one ball, there are about $n = (N-1)/L$ of them in the top layer, of which we expect about np to be red, $nq = n(1-p)$ white. Now we toss the drawn ball back in. If it was red, the probability of getting red at the next draw if we do not shake the urn, is about

$$\frac{np+1}{n+1} = p + \frac{1-p}{n} + O\Big(\frac{1}{n^2}\Big) \tag{3–121}$$

and if it is white, the probability of getting white at the next draw is about

$$\frac{n(1-p)+1}{n+1} = 1 - p + \frac{p}{n} + O\Big(\frac{1}{n^2}\Big) \tag{3–122}$$

Comparing with (3–80) we see that we could estimate ϵ and δ by

$$\epsilon \simeq q/n \ , \qquad\qquad \delta \simeq p/n \tag{3–123}$$

whereupon our magic condition (3–117) is satisfied. Of course, the argument just given is too crude to be called a derivation, but at least it indicates that there is nothing inherently unreasonable about (3–117). We leave it for the reader to speculate about what significance and use this curious fact might have, and whether it generalizes beyond the Markoffian approximation.

We have now had a first glimpse of some of the principles and pitfalls of standard sampling theory. All the results we have found will generalize greatly, and will be useful parts of our "toolbox" for the applications to follow.

COMMENTS

In most real physical experiments we are not, literally, drawing from any "urn". Nevertheless, the idea has turned out to be a useful conceptual device, and in the 250 years since Bernoulli's *Ars Conjectandi* it has appeared to scientists that many physical measurements are very much like "drawing from Nature's urn". But to some the word "urn" has gruesome connotations and in much of the literature one finds such expressions as "drawing from a population".

In a few cases, such as recording counts from a radioactive source, survey sampling, and industrial quality control testing, one is quite literally drawing from a real, finite population, and the urn analogy is particular apt. Then the probability distributions just found, and their limiting forms and generalizations noted in Appendix E, will be appropriate and useful. In some cases, such as agricultural experiments or testing the effectiveness of a new medical procedure, our credulity can be strained to the point where we see a vague resemblance to the urn problem.

But in other cases, such as flipping a coin, making repeated measurements of the temperature and wind velocity, the position of a planet, the weight of a baby, or the price of a commodity, the urn analogy seems so far–fetched as to be dangerously misleading. Yet in much of the literature one still uses urn distributions to represent the data probabilities, and tries to justify that choice by visualizing the experiment as drawing from some "hypothetical infinite population" which is entirely a figment of our imagination. Functionally, the main consequence of this is strict independence of successive draws, regardless of all other circumstances. Obviously, this is not sound reasoning, and a price must be paid eventually in erroneous conclusions.

This kind of conceptualizing often leads one to suppose that these distributions represent not just our prior state of knowledge about the data, but the *actual* long–run variability of the data in such experiments. Clearly, such a belief cannot be justified; anyone who claims to know in advance the long–run results in an experiment that has not been performed, is drawing on a vivid imagination, not on any fund of actual knowledge of the phenomenon. Indeed, if that infinite population is only imagined, then it seems that we are free to imagine any population we please.

But from a mere act of the imagination we cannot learn anything about the real world. To suppose that the resulting probability assignments have any real physical meaning is just another form of the Mind Projection Fallacy. In practice this diverts our attention to irrelevancies and away from the things that really matter (such as information about the real world that is not expressible in terms of any sampling distribution, or does not fit into the urn picture; but which is nevertheless highly cogent for the inferences we want to make). Usually, the price paid for this folly is missed opportunities; had we recognized that information, more accurate and/or more reliable inferences could have been made.

Urn – type conceptualizing is capable of dealing with only the most primitive kind of information, and really sophisticated applications require us to develop principles that go far beyond the idea of urns. But the situation is quite subtle, because as we stressed before in connection with Gödel's theorem, an erroneous argument does not necessarily lead to a wrong conclusion. In fact, as we shall find in Chapter 9, highly sophisticated calculations sometimes lead us back to urn – type distributions, for purely mathematical reasons that have nothing to do conceptually with urns or populations. The hypergeometric and binomial distributions found in this Chapter will continue to reappear, because they have a fundamental mathematical status quite independent of arguments that we used to find them here. In a similar way, exponential functions appear in all parts of analysis because of their fundamental mathematical properties, although their conceptual basis varies widely.

On the other hand, we could imagine a different problem in which we would have full confidence in urn – type reasoning leading to the binomial distribution, although it probably never arises in the real world. If we had a large supply $\{U_1 \cdots U_n\}$ of urns known to have identical contents and those contents known with certainty in advance – and then we used a fresh new urn for each draw – then we would assign $P(A) = M/N$ for every draw, strictly independently of what we know about any other draw. Such prior information would take precedence over any amount of data. If we did not know the contents (M, N) of the urns – but we knew they all had identical contents – this strict independence would be lost, because then every draw from one urn would tell us something about the contents of the other urns, although it does not physically influence them.

From this we see once again that logical dependence is in general very different from causal physical dependence. We belabor this point so much because it is not recognized at all in most expositions of probability theory, and this has led to errors, as is suggested by Exercise (3.6). In Chapter 4 we shall see a more serious error of this kind [discussion following (4–27)]. But even when one manages to avoid actual error, to restrict probability theory to problems of physical causation is to lose its most important applications. The extent of this restriction – and the magnitude of the missed opportunity – does not seem to be realized by those who are victims of this fallacy.

Indeed, most of the problems we have solved in this Chapter are not considered to be within the scope of probability theory – and do not appear at all – in those expositions which regard probability as a physical phenomenon. Such a view restricts one to a small subclass of the problems which can be dealt with usefully by probability theory as logic. For example, in the 'physical probability' theory it is not even considered legitimate to speak of the probability of an outcome at a specified trial; yet that is exactly the kind of thing about which it is necessary to reason in conducting scientific inference. The calculations of this Chapter have illustrated this many times.

In summary: in each of the applications to follow, one must consider whether the experiment is really "like" drawing from an urn; if it is not, then we must go back to first principles and apply the basic product and sum rules in the new context. This may or may not yield the urn distributions.

A Look Ahead. The probability distributions found in this Chapter are called *sampling distributions*, or *direct probabilities*, which names indicate that they are of the form; given some hypothesis H about the phenomenon being observed (in the case just studied, the contents (M, N) of the urn), what is the probability that we shall obtain some specified data D (in this case, some sequence of red and white balls)? Historically, the term "direct probability" has long had the additional connotation of reasoning from a supposed physical cause to an observable effect. But we have seen that not all sampling distributions can be so interpreted. In the present work we shall not use this term, but use 'sampling distribution' in the general sense of *reasoning from some specified hypothesis to potentially observable data*, whether the link between hypothesis and data is logical or causal.

Sampling distributions make predictions, such as the hypergeometric distribution (3–18), about potential observations (for example, the possible values and relative probabilities of different values of r). If the correct hypothesis is indeed known, then we expect the predictions to agree closely with the observations. If our hypothesis is not correct, they may be very different; then the nature of the discrepancy gives us a clue toward finding a better hypothesis. This is, very broadly stated, the basis for scientific inference. Just how wide the disagreement between prediction and observation must be in order to justify our rejecting the present hypothesis and seeking a new one, is the subject of *significance tests*. It was the need for such tests in astronomy that led Laplace and Gauss to study probability theory in the 18'th and 19'th Centuries.

Although sampling theory plays a dominant role in conventional pedagogy, in the real world such problems are an almost negligible minority. In virtually all real problems of scientific inference we are in just the opposite situation; the data D are known but the correct hypothesis H is not. Then the problem facing the scientist is of the inverse type; given the data D, what is the probability that some specified hypothesis H is true? Exercise (3.3) above was a simple introduction to this kind of problem. Indeed, the scientists' motivation for collecting data is usually to enable him to learn something about the phenomenon, in this way.

Therefore, in the present work our attention will be directed almost exclusively to the methods for solving the inverse problem. This does not mean that we do not calculate sampling distributions; we need to do this constantly and it may be a major part of our computational job. But it does mean that for us the finding of a sampling distribution is almost never an end in itself.

Although the basic rules of probability theory solve such inverse problems just as readily as sampling problems, they have appeared quite different conceptually to many writers. A new feature seems present, because it is obvious that the question: "What do you know about the hypothesis H

after seeing the data D?" cannot have any defensible answer unless we take into account: "What did you know about H before seeing D?" But this matter of previous knowledge did not figure in any of our sampling theory calculations. When we asked: "What do you know about the data given the contents (M,N) of the urn?" we did not seem to consider: "What did you know about the data before you knew (M,N)?"

This apparent dissymmetry, it will turn out, is more apparent than real; it arises mostly from some habits of notation that we have slipped into, which obscure the basic unity of all inference. But we shall need to understand this very well before we can use probability theory effectively for hypothesis tests and their special cases, significance tests. In the next Chapter we turn to this problem.