Lecture 17

PROBABILITY AND FREQUENCY IN EXCHANGEABLE SEQUENCES

We are now in a position to say quite a bit more about connections between probability and frequency. These are of two main types: (a) given an observed frequency in a random experiment, to convert this information into a probability assignment, and (b) given a probability assignment, to predict the frequency with which some condition will be realized. We have seen, in Lectures 10 and 12, how the principles of maximum entropy and transformation groups lead to probability assignments which, if the quantity of interest happens to be the result of some "random experiment," correspond automatically to predicted frequencies, and thus solve problem (b) in some situations.

The rule of succession gives us the solution to problem (a) in a wide class of problems; if we have observed whether A was true in a very large number of trials, and the only knowledge we have about A is the result of this random experiment, and the constancy of the "causal mechanism," then it says that the probability we should assign to A at the next trial becomes practically equal to the observed frequency. Now, in fact, this is exactly what people who define probability in terms of frequency do; one postulates the existence of an unknown "absolute" probability, whose numerical value is to be found by performing random experiments. Of course, you must perform a very large number of experiments. Then the observed frequency of A is taken as the estimate of the probability. As we saw in Lecture 15, even the +1 and +2 in Laplace's formula turn up when the "frequentist" refines his

methods by taking the center of a confidence interval. So, I don't see how even the most ardent advocate of the frequency theory of probability can damn the rule of succession without thereby damning his own procedure; after all polemics, there remains the simple fact that in his own procedure, he is doing exactly what Laplace's rule of succession tells him to do. Indeed, to define probability in terms of frequency is equivalent to saying that the rule of succession is the only rule which can be used for converting observational data into probability assignments.

17.1. Prediction of Frequencies.

Now let's consider problem (b) in this situation; to reason from a probability to a frequency. This is simply a problem of parameter estimation, not different in principle from any other. Suppose that instead of asking for the probability that A will be true in the next trial, we wish to infer something about the relative frequency of A in an indefinitely large number of trials, on the basis of the evidence N_n. We must take the limit of Equation (16-22) as $M \to \infty$, $m \to \infty$, in such a way that $(m/M) \to f$. Introducing the proposition

$$A_f \equiv \text{"The frequency of A true in an indefinitely large number of trials is f,"}$$

we find in the limit that the probability density of A_f, given N_n, is

$$(A_f|N_n) = \frac{(N+1)!}{n!\ (N-n)!} f^n (1-f)^{N-n} \quad , \tag{17-1}$$

which is the same as our $(A_p|N_n)$ in (16-20), with f numerically equal to p. According to (17-1) the most probable frequency is equal to (n/N), the observed frequency in the past. But we have noted before that in parameter estimation (if you object to my calling f a "parameter," then let's just call it "prediction"), the most probable value is usually a poorer estimate than the mean value in the small sample case, where they can be appreciably

different. The mean value estimate of the frequency is

$$\overline{f} = \int_0^1 f(A_f|N_n)\, df = \frac{n+1}{N+2} \qquad (17\text{-}2)$$

i.e., just the same as the value of $(A|N_n)$ given by Laplace's rule of succession. Thus, we can interpret the rule in either way; the probability which Laplace's theory assigns to A at a single trial is numerically equal to the estimate of frequency which minimizes the expected square of the error. You see how nicely this corresponds with the relation between probability and frequency which we found in the maximum-entropy and transformation group arguments.

Note also that the distribution $(A_f|N_n)$ is quite broad for small N, confirming our expectation that no reliable predictions should be possible in this case. As a numerical example, if A has been observed true once in two trials, then $\overline{f} = (A|N_n) = 1/2$; but according to (17-1) it is still an even bet that the true frequency f lies outside the interval $0.326 < f < 0.674$. With no evidence at all ($N = n = 0$), it would be an even bet that f lies outside the interval $0.25 < f < 0.75$. More generally, the variance of (17-1) is

$$\mathrm{var}(A_f|N_n) = \overline{f^2} - \overline{f}^2 = \overline{f}(1-\overline{f})/(N+3) \qquad (17\text{-}3)$$

so that the expected error in the estimate (17-2) decreases like $N^{-1/2}$. More detailed conclusions about the reliability of predictions, which we could make from (17-2) are for all practical purposes identical with those the statistician would make by the method of confidence intervals.

All these results hold also for the generalized rule of succession. Taking the limit of (16-38) as $M \to \infty$, $m_i/M \to f_i$, we find the joint probability distribution for A_i to occur with frequency f_i to be

$$(f_1 \ldots f_k | n_1 \ldots n_k)\ df_1 \ldots df_k$$

$$= \frac{(n+k-1)!}{n_1! \ldots n_k!}\ (f_1^{n_1} \ldots f_k^{n_k})\ \delta(f_1+\ldots+f_k-1)\ df_1 \ldots df_k \qquad (17\text{-}4)$$

The probability that the frequency f_1 will be in the range df_1 is found by integrating (17-4) over all values of $f_2 \ldots f_k$ compatible with $f_i \geq 0$, $(f_2+\ldots+f_k) = 1 - f_1$. This can be carried out by application of Laplace transforms in a well known way, and the result is

$$(f_1 | n_1 \ldots n_k)\ df_1 = \frac{(N+K-1)!}{n_1!\ (N-n_1+K-2)!}\ f_1^{n_1}(1-f_1)^{N-n_1+K-2}\ df_1 \qquad (17\text{-}5)$$

from which we find the most probable and mean value estimates of f_1 to be

$$(\hat{f}_1) = \frac{n_1}{N+K-2} \qquad (17\text{-}6)$$

$$\overline{f}_1 = \frac{n_1+1}{N+K} \quad , \quad \text{compare (16-39)} \qquad (17\text{-}7)$$

Another interesting result is found by taking the limit of $(M_m | A_p)$ in (16-17) as $M \to \infty$, $(m/M) \to f$. We easily find

$$(A_f | A_p) = \delta(f-p) \qquad (17\text{-}8)$$

Likewise, taking the limit of $(A_p | N_n)$ in (16-20) as $N \to \infty$, we find

$$(A_p | A_f) = \delta(p-f) \qquad (17\text{-}9)$$

which also follows from (17-8) by application of Bayes' theorem. Therefore, if B is any proposition, we have from our standard argument,

$$(B | A_f) = \int_0^1 (BA_p | A_f)\ dp = \int_0^1 (B | A_p A_f)(A_p | A_f)\ dp$$

$$= \int_0^1 (B | A_p)\ \delta(p-f)\ dp \quad . \qquad (17\text{-}10)$$

In the last step we used the property (16-1) that A_p automatically neutralizes

any other statement about A. Thus, if f and p are numerically equal, we have $(B|A_p) = (B|A_f)$; A_p and A_f are equivalent statements in their implication for plausible reasoning.

To verify this equivalence in one case, note that in the limit $N \to \infty$, $(n/N) \to f$, $(M_m|N_n)$ in Equation (16-22) reduces to the binomial distribution $(M_m|A_p)$ as given by (16-17). The generalized formula (16-), in the corresponding limit, goes into the multinomial distribution,

$$(m_1 \ldots m_k | f_1 \ldots f_k) = \frac{m!}{m_1! \ldots m_k!} f_1^{m_1} \ldots f_k^{m_k} \quad . \qquad (17\text{-}11)$$

This equivalence shows why it is so easy to confuse the notions of probability and frequency, and why in many problems this confusion does no harm. Whenever the available information consists of observed frequencies in a large sample, and constancy of the "causal mechanism," Laplace's theory becomes mathematically equivalent to the frequency theory. Most of the "classical" problems of statistics (life insurance, etc.) are of just this type; and as long as one works only on such problems, all is well. The harm arises when we consider more general problems.

Today, physics and engineering offer many important applications for probability theory in which there is an absolutely essential part of the evidence which cannot be stated in terms of frequencies, and/or the quantities about which we need plausible inference have nothing to do with frequencies. The axiom (probability) $\equiv$ (frequency), if applied consistently, would prevent us from using probability theory in these problems.

17.2. One-Dimensional Neutron Multiplication.

Our discussion so far has been rather abstract; perhaps too much so. In order to make amends for this, I would like to show you a specific physical problem where these equations apply. This was first described in a short

note by Bellman, Kalaba, and Wing (1957) and further developed in the recent book of Wing (1962). Neutrons are traveling in fissionable material, and we want to estimate how many new neutrons will be produced in the long run in consequence of one incident trigger neutron. In order to have a tractable mathematical problem, we make some drastic simplifying assumptions:

(a) the neutrons travel only in the ±x-direction, at a constant velocity.

(b) each time a neutron, traveling either to the right or the left, initiates a fission reaction, the result is exactly two neutrons, one traveling to the right, one to the left. The net result is therefore that any neutron will from time to time emit a progeny neutron traveling in the opposite direction.

(c) the progeny neutrons are immediately able to produce still more progeny in the same manner.

We fire a single trigger neutron into a thickness x of fissionable material from the left, and the problem is to predict the number of neutrons that will emerge from the left and from the right, over all time, as a consequence. At least, that is what we would <u>like</u> to calculate. But of course, the number of emerging neutrons is not <u>determined</u> by any of the given data, and so the best we can do is to calculate the <u>probability</u> that exactly n neutrons will be transmitted or reflected. I want to make a detailed comparison of the Laplace theory and the frequency theory of probability, as applied to the initial formulation of this problem. I am concerned mainly with the underlying rationale by which we relate probability theory to the physical model.

Many proponents of the frequency theory berate the Laplace theory on purely philosophical grounds that have nothing to do with its success or failure in applications. There is a more defensible position, held by some, who recognize that the present state of affairs gives them no reason for smugness,

and a good reason for caution. While they believe that at present the frequency theory is superior, they also say, as one of my correspondents did to me, "I will most cheerfully renounce the frequency theory for any theory that yields me a better understanding and a more efficient formalism." The trouble is that the current statistical literature gives us no opportunity to see the Laplace theory in actual use so that valid comparisons could be made; and that is the situation I am trying to correct here.

First, let us formulate the problem as it would be done on the frequency theory. Here is the way the "frequentist" would reason:

"The exeerimentalists have measured for us the relative frequency $p = a\Delta$ of fission in a very small thickness Δ of this material. This means that they have fired N trigger neutrons at a thin film of thickness Δ, and observed fission in n cases. Since N is finite, we cannot find the exact value of p from this, but it is approximately equal to the observed frequency (n/N). More precisely, we can find confidence limits for p. In similar situations, we can expect that about k per cent of the time, the limits (Cramér, 1946; p. 515)

$$\frac{N}{N+\lambda^2}\left[\frac{2n+\lambda^2}{2N} \pm \lambda\sqrt{\frac{n(N-n)}{N^3}+\frac{\lambda^2}{4N^2}}\right] \tag{17-12}$$

will include the true value of p, where λ is the (100 - k) per cent value of a normal deviate. For example, with $\lambda = \sqrt{2}$, the range

$$\frac{n+1}{N+2} \pm \frac{N}{N+2}\sqrt{\frac{2n(N-n)}{N^3}+\frac{1}{N^2}} \approx \frac{n+1}{N+2} \pm \sqrt{\frac{2n(N-n)}{N^3}} \tag{17-13}$$

will cover the correct p in about 84 per cent of similar cases. [Again, there's that +1 and +2 of Laplace's rule of succession!] In general, the connection between λ and k is given by

$$\frac{1}{\sqrt{2\pi}}\int_{-\lambda}^{\lambda} e^{-\frac{x^2}{2}}\,dx = \frac{k}{100}$$

Equation (17-12) is an approximation valid when the numbers n and (N-n) are sufficiently large; the exact confidence limits are difficult to express analytically, and for small N one should consult the graphs of Pearson and Clopper (1934). The number p is, of course, a definite, but imperfectly known, physical constant characteristic of the fissionable material.

"Now in order to calculate the relative frequency with which n neutrons will be reflected from a thickness x of this material, we have to make some additional assumptions. We assume that the probability of fission per unit length is always the same for each neutron independently of its history. Due to the complexity of the causes operating, it seems reasonable to assume this; but the real test of whether it is a valid assumption can come only from comparison of the final results of our calculation with experiment. This assumption means that the probabilities of fission in successive slabs of thickness Δ are independent so that, for example, the probability that an incident neutron will undergo fission in the second slab of thickness Δ, but not in the first, is the product $p(1-p)$.

"At this point we turn to the mathematics and solve the problem by any one of several possible techniques, emerging with the relative frequencies $p_n(x)$, $q_n(x)$ for reflection or transmission of n neutrons, respectively. [Actually, the analytical solution has not yet been found, but the book of Wing (1962) gives the results of numerical integration, which is equally good for orr purposes.]

"We now compare these predictions with experiment. When the first trigger neutron is fired into the thickness x, we observe r_1 neutrons reflected and t_1 neutrons transmitted. This datum does not in any way affect the assignments $p_n(x)$, $q_n(x)$, since the latter have no meaning in terms of a single experiment, but are predictions only of limiting frequencies for an indefinitely large number of experiments. We therefore must repeat

the experiment many times, and record the numbers r_i, t_i for each experiment. If we find that the frequency of cases for which $r_i = n$ tends sufficiently close to $p_n(x)$ ['sufficiently close' being determined by certain significance tests such as Chi-squared], then we conclude that the theory is satisfactory; or at least that it is not rejected by the data. If, however, the observed frequencies show a wide departure from $p_n(x)$, then we know that there is something wrong with our initial set of assumptions.

"Now, of course, the theory is either right or wrong. If it is wrong, then in principle the entire theory is demolished, and we have to start all over again, trying to find the right theory. In practice, it may happen that only one minor feature of the theory has to be changed, so that most of the old calculations will still be useful in the new theory."

* *

Now let's state this same problem in terms of Laplace's theory. We regard it simply as an exercise in plausible reasoning, in which we make the best possible guesses as to the outcome of a single experiment, or of any finite number of them. We are not concerned with the prediction, or even the existence, of limiting frequencies; because any assertion about the outcome of an impossible experiment is obviously an empty statement, and cannot be relevant to any application. We reason as follows:

The experimentalists have provided us with the evidence N_n, by firing N neutrons at a thin film of thickness Δ, and observing fission in n cases. Since by hypothesis the only prior knowledge was that a neutron either will or will not undergo fission, we have just the situation where Laplace's rule of succession applies and the probability, on this evidence, of fission for the (N+1)'th neutron in thickness Δ, is

$$p \equiv (F_{N+1}|N_n) = \frac{n+1}{N+2} \tag{17-14}$$

where

$F_m \equiv$ "the m'th neutron will undergo fission."

Whether N is large or small, the question of the "accuracy" of this probability does not arise--it is exact by definition. Of course, we will prefer to have as large a value of N as possible, since this increases the weight of the evidence N_n and makes the probability p, not more <u>accurate</u>, but more <u>stable</u>. The probability p is manifestly <u>not</u> a physical property of the fissionable material, but is only a means of describing our state of knowledge about it, on the basis of the evidence N_n. For, if the preliminary experiment had yielded a different result N_n', then we would of course assign a different probability p'; but the properties of the fissionable material would remain the same.

We now fire a neutron at a thickness $x = M\Delta$. Define the propositions,

$F^n \equiv$ "The neutron will cause fission in the n'th slab of thickness Δ."

$f^n \equiv$ "The neutron will not cause fission in the n'th slab."

The probability of fission in slab 1 is then

$$p = (F^1|N_n) = \frac{n+1}{N+2} \quad . \tag{17-15}$$

But now the probability that fission will occur in the second but not the first slab, is <u>not</u> p(1-p) as in the first treatment. At this point we see one of the fundamental differences between the theories. From our Rule 1, we have

$$(F^2f^1|N_n) = (F^2|f^1N_n)(f^1|N_n) = \frac{n+1}{N+2}\left[1 - \frac{n+1}{N+2}\right]$$

$$= \frac{(n+1)(N-n+1)}{(N+2)(N+3)} \tag{17-16}$$

The difference is that in calculating the probability $(F^2|f^1N_n)$, we must take into account the evidence f^1, that a neutron has passed through one more thickness Δ without fission. This amounts to one more experiment in

addition to that leading to N_n. The evidence f^1 is fully as cogent as N_n, and it would be clearly inconsistent to take one into account and ignore the other. Continuing in this way, we find that the probability that the incident neutron will emit exactly m first-generation progeny in passing through thickness $M\Delta$ is just the expression

$$(M_m|N_n) = \binom{M}{m} \frac{(n+m)!\,(N+1)!\,(N+M-n-m)!}{n!\,(N-n)!\,(N+M+1)!} \tag{17-17}$$

which we have derived before, Eq. (16-22). Now if N is not a very large number, this may differ appreciably from the value

$$(M_m|A_p) = \binom{M}{m} p^m(1-p)^{M-m} \tag{17-18}$$

which one obtains in the frequency approach. However, note again that as the weight of the evidence N_n increases, we find $(A_{p'}|N_n) \to \delta(p' - \frac{n}{N})$, and

$$(M_m|N_n) \to (M_m|A_p)$$

in the limit $N \to \infty$, $(n/N) \to p$. The difference in the two results is negligible whenever $N \gg M$; i.e. when the weight of the evidence N_n greatly exceeds that of M_m. Now let's study the difference between (17-17) and (17-18) more closely. From (17-17) we have for the mean value estimate of m, on the Laplace theory,

$$\overline{m} = M\,\frac{n+1}{N+2} \tag{17-19}$$

To state the accuracy of this estimate, we can calculate the variance of the distribution (17-17). This is most easily done by using the representation (16-21):

$$\begin{aligned}\overline{m^2} &= \sum_{m=0}^{M} m^2 \int_0^1 (M_m|A_p)(A_p|N_n)\,dp \\ &= \frac{(N+1)!}{n!\,(N-n)!}\int_0^1 [Mp + M(M-1)p^2]\,p^n(1-p)^{N-n}\,dp \\ &= M\,\frac{n+1}{N+2} + M(M-1)\,\frac{(n+1)(n+2)}{(N+2)(N+3)}\end{aligned} \tag{17-20}$$

which gives the variance

$$V = \overline{m^2} - \overline{m}^2 = \frac{N+M+2}{N+3} M \frac{n+1}{N+2}\left[1 - \frac{n+1}{N+2}\right] \tag{17-21}$$

while, from (17-18), the frequency theory gives

$$\overline{m}_o = Mp \tag{17-22}$$

$$V_o = (\overline{m^2} - \overline{m}^2)_o = Mp(1-p) \tag{17-23}$$

If the frequentist takes the center of the confidence interval (17-13) as his "best" estimate of p, then he will take p = (n+1)/(N+2) in these equations. So, we both obtain the same estimate, but the variance (17-21) is greater by the amount

$$V - V_o = \frac{M-1}{N+3} Mp(1-p) \tag{17-24}$$

Why this difference? Why is it that the Laplace theory seems to determine the value of m less precisely than the frequency theory? Well, appearances are deceiving here. The fact is that the Laplace theory determines the value of m <u>more</u> precisely than the frequency theory; the variance (17-23) is not the entire measure of the uncertainty as to m on the frequency theory, because there is still the uncertainty as to the "true" value of p. According to (17-23), p is uncertain by about $\pm\sqrt{2p(1-p)/N}$, so the mean value (17-22) is uncertain by about

$$\pm M\sqrt{\frac{2p(1-p)}{N}} \tag{17-25}$$

in addition to the uncertainty represented by (17-23). If we suppose that the uncertainties (17-23) and (17-25) are independent, the total mean square uncertainty as to the value of m on the frequency theory would be represented by the sum of (17-23) and

$$M^2 \frac{2p(1-p)}{N} \tag{17-26}$$

which more than wipes out the difference (17-24). The factor 2 in (17-26)

would of course be changed somewhat by adopting a different confidence level; but no reasonable choice can change it very much.

In the frequency theory, the two uncertainties (17-23), (17-26) appear as entirely separate effects which are determined by applying two different principles; one by conventional probability theory, the other by confidence intervals. In the Laplace theory no such distinction exists; both are given automatically by a single calculation. We found exactly this same situation back in our particle-counter problem [Lecture 9, Sec. 9.3.], when we compared our robot's procedure with that of the orthodox statistician.

The mechanism by which the Laplace theory is able to do this is very interesting. It is just the difference already noted; in the derivation of (17-17) we are continually taking into account additional evidence accumulated in the new experiment, such as f^1 in (17-16). In the frequency theory, the uncertainty (17-25) in p arises because only a finite amount of data was provided by the preliminary experiment given N_n. It is just for that reason that the new evidence, such as f^1, is still relevant. In thus giving a consistent treatment of all the evidence, the Laplace theory automatically includes the effect of the finiteness of the preliminary data, which the frequency theory is able to do only crudely by the introduction of confidence intervals. In the Laplace theory there is no need to decide on any arbitrary "confidence level" because probability theory, when consistently applied to the whole problem, already tells us what weight should be given to the preliminary data N_n.

What we get in return for this is not merely a more unified treatment; in yielding a smaller net uncertainty in m, the Laplace theory shows that the two sources of uncertainty (17-23) and (17-26) of the frequency theory are not independent; they have a small negative correlation, so that they tend to compensate each other. That is the reason for Laplace's smaller

probable error. If you think about this very hard, you will be able to see intuitively why this negative correlation has to be there--I won't deprive you of the pleasure of figuring it out for yourself. All this subtlety is completely lost in the frequency theory.

"But," someone will object, "you are ignoring a very practical consideration which was the original reason for introducing confidence intervals. While I grant that in principle it is better to treat the whole problem in a single calculation, in practice we usually have to break it up into two different ones. After all, the preliminary data N_n was obtained by one group of people, who had to communicate their results to another group, who then carried out the second calculation applying this data. It is a practical necessity that the first group be able to state their conclusions in a way that tells honestly what they found, and how reliable it was. Their data can also be used in many other ways than in your second calculation, and the introduction of confidence intervals thus filled a very important practical need for communication between different workers."

Of course, if you have followed everything in these lectures so far, you know the answer to this. The memory storage problem was our original point of departure, and the problem just discussed is a specific example of just what I pointed out more abstractly in Eq. (16-15). You see from (16-21). and also in our derivation of (17-21), that the only property of the preliminary data which we needed in order to analyze the whole problem was the A_p-distribution $(A_p|N_n)$ that resulted from the preliminary experiment. The principle of confidence intervals was introduced to fill a very practical need. But there was no need to introduce any new principle for this purpose; it is already contained in probability theory, which shows that the exact way of communicating what you have learned is not by specifying confidence intervals, but by specifying your final A_p-distribution.

As a further point of comparison, note that in the Laplace theory there was no need to introduce any "statistical assumption" about independence of events in successive slabs of thickness Δ. In fact, the theory told us, as in Eq. (17-16), that these probabilities are not independent when we have only a finite amount of preliminary data; and it was just this fact that enabled the Laplace theory to take account of the uncertainty which the frequency theory describes by means of confidence intervals.

Now this brings up a very fundamental point about probability theory, which the frequency theory fails to recognize; but which is essential for applications to both communication theory and statistical mechanics, as I will show in later lectures. What do we mean by saying that two events are "independent?"

In the frequency theory, the only kind of independence recognized is causal independence; i.e. the fact that one event occurred does not in itself exert any physical influence on the occurrence of the other. Thus, in the coin-tossing example that I discussed in Lecture 16, the fact that the coin comes up heads on one toss, of course, doesn't physically affect the result of the next toss, and so on the frequency theory one would call the coin-tossing experiment a typical case of "independent repetitions of a random experiment;" the probability of a heads at both tosses must be the product of the separate probabilities . But then, you lose any way of describing the difference between the reasoning of Mr. A and Mr. B in that example!

In Laplace's theory, "independence" means something entirely different, which we see from a glance at our Rule 1: $(AB|C) = (B|C)(A|BC)$. Independence means that $(A|BC) = (A|C)$; i.e. knowledge that B is true does not affect the probability we assign to A. Thus, independence means not mere causal independence, but logical independence. Even though heads at one toss does not physically predispose the coin to give heads at the next, the knowledge

that we got heads may have a very great influence on our predictions as to the next toss.

The importance of this is that the various limit theorems, which I'll say more about later, require independence in their derivations. Consequently, even though there may be strict causal independence, if there is not also logical independence, these limit theorems will not hold. Writers of the frequency school of thought, who deny that probability theory has anything to do with inductive reasoning, recognize the existence only of causal connections, and as a consequence, they have long been applying these limit theorems to physical and communication processes where, I claim, they are incorrect and completely misleading. This was noted long ago by Keynes (1921), who stressed exactly this same point.

I think these comparisons make it very clear that, at least in this kind of problem, the Laplace theory does provide the "better understanding and more efficient formalism" that my colleague asked for.

17.3. The de Finetti Theorem.

So far we have considered the notion of an A_p-distribution and derived a certain class of probability distributions from it, under the restriction that the same A_p-distribution is to be used for all trials. Intuitively, this means that we have assumed the underlying "mechanism" as constant, but unknown. It is clear that this is a very restrictive assumption, and the question arises, how general is the class of probability functions that we can obtain in this way? In order to state the problem clearly, let us define

$$x_n \equiv \begin{cases} 1, \text{ if A is true on the n'th trial} \\ 0, \text{ if A is false on the n'th trial} \end{cases}$$

Then a state of knowledge about N trials is described in the most general

way by a probability function $p(x_1 \ldots x_N)$ which could, in principle, be defined arbitrarily (except for normalization) at each of the 2^N points.

We now ask; what is a necessary and sufficient condition on $p(x_1 \ldots x_N)$ for it to be derivable from an A_p-distribution? What test could we apply to a given distribution $p(x_1 \ldots x_N)$ to tell whether it is included in our theory as given above? A necessary condition is clear from our previous equations; any distribution obtainable in the way we have derived them necessarily has the property that the probability that A is true in n specified trials, and false in the remaining (N-n) trials, depends only on the numbers n and N; i.e., not on which trials in $1 \leq n \leq N$ were specified. If this is so, we say that $p(x_1 \ldots x_N)$ defines an exchangeable sequence.

An important theorem of de Finetti (1937) asserts that the converse is also true: any exchangeable probability function $p(x_1 \ldots x_N)$ can be generated by an A_p-distribution. Thus there is a function $(A_p|X) = g(p)$ such that $g(p) \geq 0$, $\int_0^1 g(p)\, dp = 1$, and the probability that in N trials A is true in n specified trials and false in the remaining (N-n), is given by

$$P_N(n) = \int_0^1 p^n (1-p)^{N-n}\, g(p)\, dp \tag{17-27}$$

This can be proved as follows. Note that $p^n(1-p)^{N-n}$ is a polynomial of degree N:

$$p^n(1-p)^{N-n} = p^n \sum_{m=0}^{N-n} \binom{N-n}{m} (-p)^m = \sum_{k=0}^{N} \alpha_k(N,n)\, p^k \tag{17-28}$$

which defines $\alpha_k(N,n)$. Therefore, if (17-27) holds, we would have

$$P_N(n) = \sum_{k=0}^{N} \alpha_k(N,n)\, \beta_k \tag{17-29}$$

where

$$\beta_n = \int_0^1 p^n\, g(p)\, dp \tag{17-30}$$

is the n'th moment of g(p). Thus, specifying $\beta_0, \beta_1, \beta_2, \ldots, \beta_N$ is equivalent to specifying all the $P_N(n)$ for $n = 0, 1, 2, \ldots, N$. Conversely, for given N,

specifying $P_N(n)$, $0 \le n \le N$, is equivalent to specifying $\{\beta_0 \ldots \beta_N\}$. In fact, β_N is the probability that $x_1 = x_2 = \ldots = x_N = 1$, regardless of what happens in later trials, and its relation to $P_N(n)$ can be established directly without reference to any function $g(p)$.

So, the problem reduces to this: if the numbers $\beta_0, \beta_1, \beta_2, \ldots$ are specified, under what conditions does a function $g(p) \ge 0$ exist such that (17-30) holds? This is just the well-known Hausdorff moment problem, whose solution can be found many places; for example in the book of Widder (1941; Chap. 3). Translated into our notation, the main theorem is this: A necessary and sufficient condition that a function $g(p) \ge 0$ exists satisfying (17-30) [and therefore also (17-27)] is that there exist a number B such that

$$\sum_{n=0}^{N} \binom{N}{n} P_N(n) \le B, \qquad N = 0, 1, 2, \ldots. \tag{17-31}$$

But, from the interpretation of $P_N(n)$ as probabilities, we see that the equality sign always holds in (17-31) with $B = 1$, and the proof is completed.

Here is another way of looking at it, which might be made into a proof with a little more work, and perhaps discloses more clearly the intuitive reason for the de Finetti theorem, as well as showing immediately just how much we have said about $g(p)$ when we specify the $P_N(n)$. Imagine $g(p)$ expanded in the form

$$g(p) = \sum_{n=0}^{\infty} a_n \phi_n(p) \tag{17-32}$$

where $\phi_n(p)$ are the complete orthonormal set of polynomials in $0 \le p \le 1$, essentially the Legendre functions:

$$\phi_n(p) = \frac{\sqrt{2n+1}}{n!} \frac{d^n}{dp^n} [p(1-p)]^n$$

$$= (-)^n \sqrt{2n+1}\, P_n(2p-1) \quad . \tag{17-33}$$

$\phi_n(p)$ is a polynomial of degree n, and satisfies

$$\int_0^1 \phi_m(p)\ \phi_n(p)\ dp = \delta_{mn} \qquad (17\text{-}34)$$

If we substitute (17-34) into (17-27), only a finite number of terms will survive, because $\phi_k(p)$ is orthogonal to all polynomials of degree $N < k$. Then, it is easily seen that for given N, specifying the values of $P_N(n)$, $0 \leq n \leq N$, is equivalent to specifying the first (N+1) expansion coefficients $\{a_0, a_1, a_2, \ldots, a_N\}$. Thus, as $N \to \infty$, a function $g(p)$, defined by (17-32), becomes uniquely determined to the same extent that a fourier series uniquely determines its generating function; i.e., "almost everywhere." The main trouble with this argument is that the condition $g(p) \geq 0$ is not so easily established from (17-32).

The de Finetti theorem is very important to us because it shows that the connections between probability and frequency which we have found in this lecture hold for a fairly wide class of probability functions $p(x_1 \ldots x_N)$, namely the class of all exchangeable sequences. These results, of course, generalize immediately to the case where there are more than two possible outcomes at each trial.

Possibly even more important, however, is the light which the de Finetti theorem sheds on one of the oldest controversies in probability theory--Laplace's first derivation of the rule of succession. The idea of an A_p-distribution is not, needless to say, my own invention. The way I have introduced it here is only my attempt to translate into modern language what I think Laplace was trying to say in that famous passage, "When the probability of a simple event is unknown, we may suppose all possible values of this probability between 0 and 1 as equally likely." This statement, which I interpret as saying that with no prior evidence, $(A_p|X) = \text{const.}$, has been rejected as utter nonsense by virtually everyone who has written on probability theory in this century. And, of course, on any frequency definition

of probability, Laplace's stattment could have no justification at all. But on any theory it is conceptually difficult, since it seems to involve the idea of a "probability of a probability," and the use of an A_p-distribution in calculations has been largely avoided since the time of Laplace.

The de Finetti theorem puts some much more solid ground under these methods. Independently of all conceptual problems, it is a mathematical theorem that whenever you talk about a situation where the probability of a certain sequence of results depends only on the number of successes, not on the particular trials at which they occur, all your probability distributions can be generated from a single function g(p), in just the way we have done here. The use of this generating function is, moreover, a very powerful technique mathematically, as you will quickly discover if you try to repeat some of the above derivations [for example, Equation (16-22)] without using an A_p-distribution. So, it doesn't matter what you or I might think about the A_p-distribution conceptually; its validity as a mathematical tool for dealing with exchangeable sequences is a proven fact, standing beyond the reach of mere philosophical objections.