CHAPTER VI

INTERPRETATION AFTER AN EXPERIMENT

In order to see how we will be able to interpret our results, we must look closely at the questions that our experiment is to ask, and how we are going to conduct it, with special attention to problems of variation and risks of bias. At this stage we are not concerned with "interpretation" in the sense of a profound explanation of the observed phenomena, and we will use "causal" in the everyday (nonmetaphysical) sense of the term.

This chapter contains much discussion of "statistical significance tests," and one might well ask: "Why not describe the tests first and talk about them afterward?" The reasons for not adopting that sequence are threefold:

1. Underlying all tests of statistical "significance" and all estimation of population values from samples, there are certain principles, derived from our experience of random processes ("chance" in the strict technical sense of the word). Unless we grasp these principles, and unless we know what the tests and estimates can and cannot tell us about the real world, the techniques are dangerous.

Before becoming involved with the special features and arithmetic of any particular technique we can, without great difficulty, start to grasp the principles by reference to such phenomena as card shuffling and the sampling of disks in a barrel. Such phenomena are not only our chief sources of knowledge of chance, but are used in modern experimentation in the real world. By starting in this way we can avoid the need for repetition of principles, and of warning against dangers, with each individual technique.

2. Many of us, having learned a few techniques, illustrated by simple examples in a textbook or classroom, have applied them to complex data, or to bits of such data, without knowing what fools we were making of ourselves. If we had known more about the tests in general, we would have had a better notion of what kind of test to seek for and what to avoid. Often, we might have decided not to use a test at all. Indeed,

if this chapter helped to reduce the "incidence" of tests, but promoted simpler, more thoughtful and more systematic investigation, followed by more cautious interpretation, I would be very happy. At all events, it may help to reduce the incidence of reader-deception by tests that will doubtless continue to appear in medical journals.

3. Individual statistical tests may come and go, but random processes in Nature will always be with us, mixed with trends, systematic differences and other sources of bias. They will even remain with those workers who seem to think that they can shut out chance by closing their laboratory doors. Even if we do not contemplate applying any of the tests that the statisticians have devised, the principles underlying those tests provide a groundwork for a kind of thinking that we must do when we are planning, performing and interpreting any investigation.

Q VI – 1. What exactly is the hypothesis that we wish to test by our experiment?

This is a more specific question than Q II – 1 – What is the immediate purpose of our investigation? Let us suppose that our hypothesis is that drug A, administered to a certain type of patient under certain specified conditions, causes improvement in a higher proportion of patients than does drug B administered to the same type of patient under the same conditions. Obviously, our hypothesis would not be proved true merely by showing that in the A-patients a higher proportion had improved than in the B-patients. We must show that the observed difference would rarely occur unless A was more effective than B. That is why we arrange that the random assignment of drugs to patients shall be the only difference-causing factor besides the possible difference in drug effects. We know how often the randomization itself causes differences of various magnitudes, and if the observed difference is rare according to the standard that we have chosen, we accept it as proof of our hypothesis regarding the drugs.

The Null Hypothesis. In reality the hypothesis that we are testing – the one that gives us numerical values (numbers improved and not improved) by which we judge our observed values – is the hypothesis that there is no difference between the drug effects. Such a "no-difference" hypothesis, involving either frequencies or measurements, is commonly called a "null" hypothesis. By disproving (rejecting) the null hypothesis we prove (accept) the original hypothesis; but it is important to note that we can never prove that the null hypothesis is true. Treatments may really differ in their effects, but not enough to create in the experiment a difference that is any greater than what is often produced by randomization alone.

Null hypotheses are often illustrated by comparisons of two or more treatments or other independent variables (A, B, and so on), the dependent variable (X) being a frequency or a measurement. However, null hypotheses are equally basic in the study of relationships between measured variables, such as different amounts of a certain dietary supplement (A_1, A_2, A_3, and so on) and the amounts of growth (X_1, X_2,

X_3, and so on) in different children or groups of children during a certain period. After the experiment, we draw a dot diagram, with the A's on the horizontal axis, the X's on the vertical axis, and a dot for each child. If the null hypothesis (no AX relationship) were true, and there were no other source of variation (individual differences in amounts of growth) the dots would lie along a horizontal line; but of course such dots are always scattered, and there may be a suggestion of trend. A glance may suffice to tell us that this suggestion would be often produced by the randomization that assigned the amounts (doses) of supplement to the individual children; but for greater assurance we may perform a numerical test.

Hypotheses in General. The aura of dignity and importance that often surrounds the word "hypothesis" would be removed if we remembered that it is merely the Greek form of the Latin *suppositio.* It is a supposition that we express in such a form that we can test it. Hence, three general requirements should be obvious:

1. The hypothesis must be clearly defined. This often necessitates the breaking down of a complicated hypothesis into its components.
2. The consequences of the hypothesis must be fully worked out – what would happen and what would not happen, if the hypothesis were (a) true and (b) false.
3. The experiment must be so designed that its outcome can be placed alongside the deductions obtained in (2).

"Hypothesis" when loosely used can be very misleading. When a person who is not familiar with research decides to conduct an investigation he may use the word "hypothesis" as equivalent to "premise" or "axiom" or "belief"–something that he takes for granted, not something to be tested by the investigation.

Q VI – 2. By what rules are we going to insure that the causal interpretation after the experiment will take the form: "Either the randomization or the factor under test"?

The two rules given in the drug trial example under Q II – 4 can now be repeated in more general terms:

1. Assign the sampling units to the variants of the factor under test by a strictly random method.
2. During the experiment permit nothing to interfere with the effects of the randomization except the factor under test; that is, insure that the randomization is solely responsible for the biases throughout the experiment.

The main purpose of this chapter is to amplify those rules.

Q VI – 3. Is randomization necessary if we are not going to apply a statistical test?

Probably the best way to answer that question is to consider what a good experimenter does after an experiment. Even if he is going to apply an arithmetical test later, he looks at the results and tries to decide whether the numerical differences, which appear to be associated with treatment differences, could easily have been caused by something else, such as intersubject variation or measurement error or biases that he could not control. In so doing he actually applies a statistical test – the "eye test." Of course, the eye cannot detect a hidden bias, but neither can arithmetic. As Wilson (1952) has remarked, "fifty pages of higher mathematics will not salvage an experiment with a hidden bias."

The purpose of randomization is to control the biases that may result from the variability of our material, our procedures and other circumstances affecting the experiment. Therefore the question about the need for randomization is best put in another form.

Q VI – 4. What is the risk in trusting to our pre-existing knowledge of variability and to our way of conducting the experiment, instead of using randomization?

To an increasing number of experimenters this would be a pointless question. They would say: "Randomization is an integral part of our experiments." However, the question needs consideration because this attitude is by no means universal in medical laboratories.

Randomization as a Routine Procedure. In medicine we could learn from industrial laboratories an appreciation of the value of routine randomization, as in the following incident reported by Wilson (1952).

In the manufacture of a certain plastic object, hot plastic was introduced into a mold and pressed into shape. In order to ascertain the effect of duration of pressure upon the strength of the object, six batches of plastic were pressed in the same mold, the first batch for 10 seconds, the second for 20 seconds, and so on. When the strength of each object (Y) was plotted against the duration of pressure (X) the six points formed a smooth ascending curve, which was interpreted as showing a strong dependence of strength on duration of pressure.

However, the research supervisor objected to the experiment because the durations of pressure had not been arranged in random order. When the experiment was repeated with the sequence of the six durations randomized, there was no suggestion of a relationship between duration and strength. The source of the error in the first experiment was easily discovered. As successive batches of plastic were introduced, the mold became hotter, and the higher temperature was the cause of the greater strength of the objects. Time and treatment had been confounded.

Incidentally, as Wilson pointed out, even when randomization is used it is important to keep accurate records of the order in which tests are made. When the data from the second experiment were plotted with strength as Y and order as X, the same kind of relationship was found as in the first experiment, and this gave the clue to the cause.

Reasons and Excuses for Not Randomizing. It is desirable to look at

three of the reasons expressed or implied by medical laboratory workers who do not see why they should use randomization.

1. "The reasoning after my type of experiment is not statistical. I reduce my experimental error to a very small quantity and then accept as 'real' only those measurement differences that are of a different order of magnitude from my error." In certain lines of medical research akin to "pure" chemistry and "pure" physics this attitude is certainly tenable, but the statement suggests a confusion between statistical tests and statistical thinking. Many workers who claim that they do not use statistics do not feel comfortable unless they have repeated an experiment at least once, on the "off chance" that something "unusual" happened the first time.

2. "I am so well acquainted with the variability of my material and with my experimental error that any hidden bias will be unimportant. If I suspect a more serious bias I do another experiment designed so as to avoid it." This argument may be legitimate for certain workers in certain of their studies, and when they are observing by measurement rather than enumeration, but when it is adduced as a general defense of individual judgment it prompts the following three reflections:

a. It appears to imply that an experimenter knows everything important about an experimental situation except the factor that he is testing. It would seem, therefore, more appropriate to routine industrial testing than to original research, which is a probing into the vast unknown. And yet randomization has come to be recognized as vital in industrial testing, perhaps because mistakes and wastage can be measured there in dollars.

b. The dependence on previous experience prompts the question: "How can we be sure that the new situation was like the previous situation, qualitatively and quantitatively, in all relevant respects?" This does not imply that we should ignore past experience in setting up an experiment and evaluating its outcome. What we are asking is that the experiment shall itself give us information by which we can evaluate what it seems to tell us.

c. Estimates of experimental error are often not nearly as reliable as they are supposed to be. Useful but sometimes embarrassing questions are: "What exactly do you mean by the statement that your experimental error is $\pm$ 2 per cent?" "How often would you expect that value to be exceeded in a hundred experiments?" "How many observations, under what conditions, provided your estimate of error?" It would often take several hundred observations to justify the confidence that many experimenters place in an error estimate derived from a dozen or twenty observations. Even an extensive special study of error leaves unanswered the question: "Will the conditions in the experiment in which we are going to use this estimate be exactly the same as in the study that produced it?"

3. "The biases that we have not eliminated by the design and conduct of the experiment are due to chance and can therefore be allowed for by a statistical test." This still rather common faith in statistical arithmetic seems to reflect a confusion between the colloquial and technical use of

the word "chance." Experimenters are not entirely to blame, because the confusion can be traced not only to statistical "cookbooks" but to some of the fundamental writings that introduced modern statistics to experimenters (Mainland, 1960).

Randomization in Medical Laboratories. Unlike the industrial laboratory worker who tested the plastic mold, distinguished professors in the medical sciences do not have research supervisors who will tell them to do an experiment over again, introducing randomization. It would perhaps help such professors if they would open-mindedly share the experience of a biological statistician who has been called in to perform an "autopsy" on a long series of experiments. He often finds traces of havoc caused by failure to randomize when he compares data from animals "treated alike," and close together in time, with data from animals that were also "treated alike" but farther apart in time. He can seldom do more than raise doubts which may cause a wise experimenter to discard the results of months of work. The reader of a journal article is of course much less able to detect such faults than is a person who can talk to the experimenter.

Perhaps the best advertisement for randomization is the feeling that it confers on the experimenter—a freedom from worry about bias. However carefully we set up and conduct an experiment, unforeseen and unpreventable things are almost certain to occur. An instrument may break down. We repair it or replace it by another of the same kind. We may be able to apply a correction term to all readings taken by the new instrument, trying to make them equivalent to those from the first instrument; but we cannot be sure that the corrected readings are close enough to those that the first instrument would have given so that we will have no hidden bias. Even when an instrument functions as well as possible throughout an experiment it may fluctuate or drift, and although we repeatedly check it and apply correction terms we can never be quite sure that they are adequate.

The animals or other material on which we are experimenting may be different, perceptibly or imperceptibly, at different times in an experiment. The experimenter may unwittingly alter certain methods of procedure, observation or assessment during the course of the experiment.

We ought not to be careless about these and many other possibilities or fail to correct for disturbances when we can do so, because any increase of variation reduces the sensitivity of our experiment. But if we have placed our sampling units in random relationship to such events we know that we can make proper allowance for the bias that they may have introduced.

Q VI – 5. What are the risks in using the method of "alternates" instead of true randomization?

From time to time this question is still asked, because this method seems to be such an easy "unbiased" way of assigning treatments to patients in their order of admission to a clinical trial or to animals in the

order in which they are taken out of a cage—treatment A to Nos. 1, 3, 5, and so on; treatment B to Nos. 2, 4, 6, and so on. The objections to the method can be described as logical and psychological.

Logical Objections to Alternate-subject Assignment. A clinician, talking to a statistician, said: "In a drug trial I arranged that the clinic attendant would hand out the A and B drugs to alternate patients as they left the clinic. What was wrong with that as a method of randomization?" An appropriate answer would have been: "How did you prove that it was right?" As in the selection of every *n*th subject from a sample (Q III – 8), whenever anyone proposes to substitute a systematic selection method for a strictly random method the onus is on him to prove that the substitute is sufficiently equivalent to the random method for his purpose.

When numbers are taken from a roulette wheel for the construction of a table of random numbers, it requires an enormous exploration, with thousands of digits, to prove that the wheel is sufficiently unbiased. In an experiment containing a few score subjects such a proof would be impossible, even for the characteristics of the subjects that we can observe and record; and we know, moreover, that there are likely to be undetectable characteristics that will have great bearing on the subjects' reactions in the experiment.

We can distinguish two ways in which the alternate-subject assignment can interfere with the logical structure of an experiment: (1) by introducing bias, and (2) by automatic matching.

Bias Due to Alternate-subject Assignment. During an outbreak of infectious disease the severity of successive cases may in general increase during the earlier stages of an outbreak and decrease during the later stages. If during the later stages we apply treatments A and B to equal numbers of patients in the order ABAB, the B's will on the average be less severe cases than the A's.

When we reach haphazardly to take animals from cages, our selection is far from random. The lively ones that pop up may be the ones we take first, or it may be easier to catch the sluggish ones. In fact, weight trends have been demonstrated in more than one series of animals so selected. The ABAB design applied to such series invites bias. Similarly, if during a 6-hour laboratory period 6 animals are subjected to one or other of two operations, A and B, in the order ABABAB, the B's will have been subjected for an hour longer than the A's to whatever influences the waiting period under laboratory conditions may exert on them, and these influences are often far from negligible.

A grossly simplified picture, fictitious but based on an actual investigation, will help to show how unsuspected and probably undetectable rhythms may cause bias. A surgeon wished to find out whether a certain seasickness remedy would reduce the incidence or severity of postoperative vomiting. In order to avoid biased selection, he administered the drug to alternate patients.

Let us suppose that there are two surgical operating days, Mondays and Fridays, and that there is some factor that tends to make the inci-

dence or severity of vomiting greater on one of the days than on the other. Such a factor might be one of the following: the relationship of the operating day to a weekend or to a hospital visiting day, which often disturbs patients; different nurses or different attitude of the nurses; the surgeon's other activities; difference in preoperative preparation, such as maintenance of body fluid balance. Let there be on each day some most common number (modal number) of the operations used in the study: 5 on Mondays, 3 on Fridays. For simplicity we imagine these numbers constant. Allocating the drug (D) and no-drug (N) to alternate patients, we have in each week:

Monday, DNDND; Friday, NDN.

There are equal numbers of D's and N's, but if Monday tends to be the "more favorable" day the drug will be favored. In practice there might be other biases that would counteract this bias, but we have no right to assume so. More complicated rhythms, within any time period from a day to a year, can be similarly conceived.

Automatic Matching by Alternate-subject Assignment. If two subjects, taken at the same time or close together in time, are more alike than those taken farther apart, the ABAB sequence will create a series of more or less matched pairs. This could occur in the examples given above —the trend of increasing (or decreasing) severity in an outbreak of disease, and the selection of animals from cages. It could conceivably have occurred in the drug trial in which the attendant handed out drugs A and B to alternate patients as they left the clinic. Junior physicians may deal with the less difficult or less serious cases and dismiss them early in the clinic period, whereas the more complicated and more serious cases will be retained for consultation with the clinic chief, perhaps with other specialists and perhaps to receive special advice or treatment.

Now if we have in reality a series of matched pairs and then analyze the data as if we had randomly assigned the treatments through the whole group (e.g., by contingency table analysis) we are applying an inappropriate method of analysis. Each member of the A-sample will have its mate in the B-sample, and the outcome in the two samples may be much more alike than if we had created the samples by overall randomization. When we allow for this larger variation before being willing to accept the proof of an A-B difference, we are making an unnecessarily large allowance and we may miss a difference that actually exists.

The alternate-subject method of assignment does not in fact lead to any valid analysis. There is always the risk of confounding sequence and treatment. If we wish to use matched pairs in spite of their drawbacks (see under Q IV – 6) we must assign A and B within each pair strictly at random and then set up the results for analysis in terms of pairs.

Psychological Objections to Alternate-subject Assignment. One of the most notable instances of vast labor wasted by a systematic assignment of therapy is described by Truelove in *Medical Surveys and Clinical*

Trials (Witts). In a trial of anticoagulant therapy in coronary heart disease it was arranged that patients admitted to the participating hospitals on odd-numbered days of the month would receive the test treatment and those admitted on even-numbered days would not. After the trial was over it was found that there were 589 treated patients and 442 controls. This was far too big a difference to be accounted for by the slight excess of odd-numbered days in a year. It was then reported that some patients had been put into the treatment group at the request of their relatives and private physicians, and it is impossible to determine how many others may have been steered into it by arranging for their admission on the odd-numbered days. Consequently, no one can say what the results of this very large trial really mean.

When I hear a clinician's proposal to assign treatments according to whether the patient's hospital admission number is odd or even (because he cannot see how this assignment could be selectively manipulated), I am surprised at his lack of faith in man's ingenuity.

The psychological danger of the alternate-subject and similar methods of assignment lies not only in the risk of purposive steering of patients to one or the other treatment. When a conscientious physician is deciding whether a patient qualifies for admission to a trial, he is afraid of being biased by knowing which treatment will be administered. In his effort to avoid bias he may come to a different decision regarding admission from the decision he would have reached without that knowledge.

Q VI – 6. What instrument are we going to use for randomization?

Tables of Random Numbers. These tables provide the easiest and most dependable method of randomization, and they are now readily available to research workers. They can be looked upon as equivalent to thousands of thoroughly shuffled playing cards, each card bearing one digit from 0 through 9. The digits, when originally recorded in random order (by methods to be mentioned) are, as it were, in one long line, but are broken up into columns, rows and blocks to facilitate reading. The following is a small portion of Fisher and Yates' (1938-1957) six-page table which contains 15,000 digits:

26	72	39	27	67
43	00	65	98	50
16	06	10	89	20
09	65	90	77	47
65	39	07	16	29

To use the table, we select a digit anywhere in it without previously inspecting the numbers themselves, e.g., by opening at a page and placing on it anywhere a pointer such as the corner of an index card. We start at the digit touched by the pointer and proceed by consecutive digits up or down or to right or left, passing from block to block without interruption. Having come to the end of a row or column we can start at the next one on either side and proceed in the same or reverse direction.

If we are performing a series of investigations that will be connected with each other we must avoid using the same sequence of digits on two occasions. Therefore it is desirable to keep a note of the part of the table that we have used.

Examples of the Use of Random Numbers. A few simple small-sample illustrations will indicate the principles. (For complex experiment designs or very large samples, special methods published with the various sets of tables are desirable in order to prevent one's using up too many digits.) The examples can be expressed as instructions to the randomizer, who should check each step very carefully before he goes on to the next one.

Ex. 1. Assignment of treatments A and B, each treatment to 5 patients.° Take ten index cards (5″ × 3″) and write on each card a number representing one of the patients (e.g., the order in which he will be admitted to the trial). Suppose that the pointer has landed on the zero of "10" in the middle of the above table. Write on the first card "10," on the second card "89," on the third card "20," on the fourth card "09," and so on. Arrange the cards in a pile in ascending (or descending) order of the random numbers and then take off the top five cards and mark them "A"; mark the remainder "B." The systematic arrangement of the random numbers has transferred the random order to the patients.

	Random No.	Patient No.
A	09	4
	10	1
	20	3
	39	10
	47	8
B	65	9
	65	5
	77	7
	89	2
	90	6

The duplicates (two 65's) cause no disturbance because both of them have led to the assignment of treatment B. If, however, one 65 had been the fifth item and the other the sixth item it would have been necessary to decide whether patient No. 5 or patient No. 9 should receive drug A. This could have been decided by picking up two random numbers somewhere else in the table and letting their ascending order of magnitude determine the relative positions of the cards with duplicate numbers. We can do likewise with triplicates and other ties; but it is

° As Herrera (1955) has pointed out, a method that was prescribed in the first edition of this book for separation of subjects into two or more groups by random numbers did not entirely remove the risk of bias.

often best to reduce the number of such problems by using four-digit numbers for the randomization.

To resume the procedure, make a list of patients and their assigned treatments:

Patient No.	Treatment
1	A
2	B
3	A
4	A
5	B
6	B
7	B
8	A
9	B
10	A

It is useful to arrange the A's and B's in separate columns, to prevent misreading of letters, and in case carbon copies of the list should become smudged.

If the trial is to be double-blind, a copy of this list can be sent to the drug dispenser who will put the patient's numbers on the proper bottles or other containers and remove all clues to the identification of drugs. If the trial is not to be double-blind the physician and all others concerned with the patient can be kept in the dark, until treatment is actually to begin, by preparing an envelope for each patient with his number on the outside. Inside is placed a card with the patient's number and also the assigned treatment, along with paper or cards to prevent anyone from being tempted to hold the envelope up to a bright light. The envelopes are sealed and delivered to the physician or other responsible person.

NOTE. — If the number of available patients were odd, say 11, the top five (or six) cards could be marked "A" and the remainder "B"; no uncontrolled bias would result.

Ex. 2. Treatments A and B are to be applied to 10 pairs of litter mates, one treatment to each animal. Write down an identification number or letter for one member of each pair. Assign to these animals the treatments by single-digit random numbers, letting even numbers (including 0) represent treatment A and odd numbers treatment B. The other member of each pair will receive the other treatment.

Ex. 3. Ten tubes in a laboratory (or 100 x-ray films or 50 histopathologic slides) are to be arranged in random order. Write the serial number or other identification of each tube (or film or slide) on an index card and proceed as in *Ex. 1*, stopping short of division into A and B groups.

Ex. 4. There are 234 animals in stock and a random sample of 40 is

required. First, identify each animal. If they are not in single-animal cages, a temporary or permanent mark can be made on each of them. On each of 234 cards write the identification of a particular animal. Write a four-digit random number on each card, arrange the cards in ascending order of random numbers and take the first 40 cards to indicate which animals shall comprise the sample.

Ex. 5. An intersubject comparison of treatments A and B has to be completed within a certain number of months. It is difficult to predict how many subjects will become available during the period. It would be permissible to assume a certain total and randomly assign A and B to equal numbers. Even if the assumed total was not reached by the end of the period, and even if the numbers in the treatment classes were unequal, the randomization would still control the bias. However, more than the anticipated number of subjects might arrive, and the combination of two sets of observations, each randomized within itself, presents some problems. The best solution seems to be random assignment of the treatments to each subject at the very beginning of the experiment, using odd numbers for A and even numbers for B, without trying to make the samples equal in size. One can assign treatments to more than the anticipated numbers of subjects, but if one runs short of the assignments before the end of the experiment, one can supplement them by the same process.

Manufacture and Testing of Random Numbers. At present there are in common use three large sets of random numbers:

1. The table of Fisher and Yates (1938–1957) already mentioned. It was derived from the 15th to 19th digits in a twenty-figure logarithm table, the selection and arrangements of the digits being determined by two sets of playing cards.
2. The tables of Kendall and Babington-Smith (1939, 1946), containing 100,000 digits. This was produced by the use of a disk with digits 0 to 9 equally spaced around it. The disk was rotated at uniform speed in the dark, and was illuminated at irregular intervals by flashes of light from a lamp controlled by an operator working a telegraph key. Each digit thus rendered visible was recorded. (It is recommended that these tables be read horizontally because they were more thoroughly tested in that direction than vertically.)
3. *A Million Random Digits* produced by the Rand Corporation (1955) by an equipment constructed and operated on the principle of a roulette-wheel but employing an electronic random frequency pulse.

For many small investigations a sufficient supply of random numbers is to be found in *Tables for Statisticians* by Arkin and Colton (1950–1959), which reproduces the first 8000 digits of the Kendall and Babington-Smith tables.

All these tables were thoroughly tested by the application of our knowledge of what happens in random processes — for instance, the

approximately equal frequencies of the digits 0 to 9, the extent to which departures from equality occur, the intervals between digits of the same value, and various other relationships analogous to those found in card games (e.g., one of the tests applied to random numbers is called the "poker" test).

"Pure chance" – a completely random arrangement or sequence – is of course an ideal, a limit which natural phenomena and man-made instruments (like tables of random numbers) can approach but can never reach. The tables doubtless contain spots of nonrandomness, but they are the best instruments for randomization available to the experimenter, and have been far more thoroughly tested than many of his other instruments.

Q VI – 7. What sources of variation must we remember in planning the randomization?

Perhaps the most complete coverage of potential sources of variation is provided by the familiar five questions mentioned in Chapter I: *Who? What? Where? When? How?*

With these questions in mind we should try to visualize circumstances and events before and during the experiment, remembering that each sampling unit will meet the resultant (net effect) of a different combination of these variables. If the treatments to be compared are A and B, the randomization must determine whether the sampling unit that meets any particular combination of variables is an A-treated or a B-treated unit.

After contemplation of all this complexity it is consoling to remember that in a clinical trial extending over many months one simple initial randomization is often sufficient. However, many common types of medical experimentation appear less simple in design than clinical trials and some guidelines for randomization are helpful.

Q VI – 8. What are we going to randomize, and how?

Without any reference to a statistical test which we may later apply, let us imagine that we are inspecting the measurement data after an experiment that involved treatments A and B applied to different subjects. We try to see whether measurements within each treatment group agree closely with each other, and whether there is little overlap between the groups. That is, we use the intragroup variation as a kind of yardstick to measure the intergroup difference. Similarly, when we are looking at enumeration data, e.g., percentages of deaths and survivals, we look for similarity within the treatment groups and difference between them – a high percentage of deaths in the one group and a high percentage of survivals in the other.

We can trust our conclusions regarding the difference of the effects of the two treatments if we can be sure that, except for the treatments pushing the values apart, the randomization has been responsible for the presence of each value in its respective treatment group – that is, responsible for the intragroup variation that we use as our yardstick.

With this in mind we can try to formulate some general guidelines or rules for randomization.

Seven General Rules for Randomization. These seven rules are followed by examples.

1. Be sure that your experiment design is simple enough for you to comprehend and carry through.
2. Randomize all the variation that you are going to use as a yardstick.
3. Do not randomize variation that you are not going to use as a yardstick. If you do, you will lower the sensitivity of the experiment.
4. Do not use randomization to make treatment groups *alike* in their characteristics. That is not its purpose. If you wish to make them alike, do so before the randomization, by a systematic restriction or subdivision of the population to be studied, or by making experimental conditions more uniform.
5. Avoid a confused mixture of random and systematic arrangements. It will lose more than it gains, because only bits of the data will be comparable.
6. If the experiment is to answer more than one question, consider whether more than one randomization will be necessary.
7. Remember that interclass variation + intraclass variation = total variation. Therefore, if a systematic design is to be used in order to remove interclass variation, so that treatments can be tested within classes, put as much of the total variation as conveniently possible into the interclass category.

Examples of Randomization. In the following six examples no attempt is made to cover all the problems of design in the projects discussed, or to present alternative designs. The examples try to obey Rule 1, even though a more complex design might provide more information from the same size of experiment.

Ex. 1. A long-term animal experiment, e.g., comparison of foods or potential cancer-producing agents. The simplest design is to place the animals in their cages and randomize treatments throughout. Interanimal variation will include effects of cage position (light, temperature, humidity and ventilation), but the removal of cage-position effects would require a more complex design. If some or all of the cages contain more than one animal, it may be found that all the animals in certain cages have been assigned the same treatment; but we must not interfere with this randomization-effect by arranging, say, for 2 A's and 2 B's to be together in the same cage (Rule 5). Single-animal cages avoid such clustering and have other advantages also.

Ex. 2. Litter mates in an A-B long-term experiment. Each pair is a class or "block" and the sampling units are the individual animals within each pair. Our yardstick is the difference between readings on pairs, one difference from each pair. We know, for instance, that if A and B were merely letters assigned randomly and if every reading on a B-animal

were subtracted from the reading on the corresponding A-animal, we would in the long run approach 50 per cent positive and 50 per cent negative differences; and for the purposes of our finite experiment we know how often various departures from the 50:50 ratio are met in samples of various sizes.

We are not going to use as our yardstick the interlitter differences that are going to affect equally both animals of a litter. Therefore we do not arrange the litters in random order in the laboratory (Rule 3); we arrange them in any convenient order. We reduce the possible differences between litter mates (except the differences that the treatments may cause) by keeping the litter mates close together. That is, we put much of the interanimal variation, which may be due to differences in location, into the interlitter variation, because we are not going to use this in the analysis (Rule 7).

Ex. 3. An acute experiment – comparison of A and B treatment on different animals. Treatment and observation on each animal will be completed in about an hour, but the whole experiment will extend over weeks or months. It is tempting to plan for an equal number of animals (say three) randomly assigned to each treatment on each day of the experiment, but unless we feel very sure from past experience that we can fill our quota each day the plan is risky. Some statistics books show how to analyze measurement data with unequal numbers of subjects in the treatment groups in different blocks (days), but apart from the complexity of the analysis it often involves assumptions that an experimenter would question if he could understand what they were.

The simplest procedure is to randomize in advance for the whole experiment (see *Ex.* 5 under Examples of the Use of Random Numbers, p. 81). Even if this should assign the same treatment to all animals on one day we should resist the desire to tinker with the randomization (Rule 5). To avoid risk of bias through foreknowledge of the scheme of each day's work, someone who does not know the scheme can pick out the animals, or sealed envelopes can be used as in drug trials, one to be opened for each animal after it has been chosen.

Ex. 4. Intrasubject comparison of drugs A and B by cross-over design – a 3-month period on each drug. A certain measurement is taken at the beginning and end of each period, and the change in period (1) is compared with the change in period (2). It might be thought that the periods in each patient could be looked on as sampling units, analogous to litter mates or twins; but the difference, period (1) minus (2), might be positive (or negative) in all or nearly all the patients even if there were no drug difference. Therefore the experiment fits more clearly into the scheme of the null hypothesis if we consider the patients as sampling units, each presenting one measurement, a period (1) minus period (2) difference.

Then the "treatments" that we compare are the sequences, AB versus BA. Hence, we randomly assign these sequences (usually in equal numbers) to the patients. In the one sequence group, the difference (1)

minus (2) means A minus B, in the other group it means B minus A. Therefore if, after the experiment, the two sequence groups differ in their (1)-minus-(2) differences more than we are prepared to attribute to the randomization, we attribute it to the difference in drug effects. (Further consideration of this design is best postponed until analysis of data is discussed – frequency data in Chapter XII, measurement data in Chapter XIV.)

Ex. 5. After an intersubject comparison of a drug with a placebo in rheumatoid arthritis, the patients' hand films were sent to a radiologist to determine whether there was less progression of the disease, between pre- and posttreatment films inspected side by side, in the drug-treated patients than in the placebo patients. The radiologist was not to know about the individual patients' therapies, but the question arose: Should the films be arranged in random order?

The random assignment of patients in the trial had randomized the relationship between the patients' order of entry and the treatments that they received. Therefore, to answer merely the question regarding treatment and x-ray progression of the disease, the radiologist could have examined the films in the order in which the patients had entered the trial, just as did the clinicians who assessed each patient at the end of his 6-month treatment.

However, there was an opportunity to inquire also into the relationship between the clinicians' findings and the radiologist's findings (Rule 6). Therefore, trends of various kinds had to be considered. In a trial in which suitable patients are scarce, a clinician, having utilized all the immediately available patients, is prompted to call in the patients who, although qualified in terms of the plan, may be somewhat different in type or severity from the patients admitted earlier. Again, the clinician's own standards of assessment may have altered during the trial, however closely he has tried to follow the prescribed system. The radiologist may also be the victim of nonrandom trends and fluctuations, e.g., an initial warming-up period, followed by a fairly steady state, perhaps later followed by monotony or fatigue, and at the end perhaps pressure to get the survey completed.

The trend or fluctuations in patient type would be combined with the clinical observer's trend in his reports. The radiologist's trends and fluctuations might, to a greater or less extent, coincide with or run counter to those in the clinicians' reports. Such complex phenomena are difficult to eliminate by analysis of the data. Therefore the film envelopes, each containing one patient's pair of films, were arranged in random order and then numbered 1, 2, 3, and so on, to show the order in which they were to be examined. This placed the radiologist's variation in random relationship to the variation in the clinical records.

Ex. 6. Observational variation estimated by duplicate readings. Probably the chief reason why many estimates of observational variation (experimental error or reading error) are too low is that duplicate readings are made one immediately after the other. To obtain a reliable estimate, the readings (two on each specimen) should be spread inde-

pendently at random throughout the whole period of the experiment, or series of experiments, in which they are to be used (Rule 2). This can be done easily on preserved material like histologic specimens or x-ray films, but perishable specimens present difficulties. In the analysis of serum the difficulty can often be overcome by freezing and storing a sample of each specimen; but to avoid the effects of thawing and refreezing it may be desirable to store two samples of each specimen.

Q VI – 9. Are we going to examine our treatment assignments in order to see if the randomization has "worked well"?

Sometimes in reports on clinical trials the reader's attention is called to the evenness with which the randomization has distributed characteristics such as sex and age between the treatment groups. This may help to increase confidence in a technique that is still unfamiliar to many readers, but it is a somewhat curious ground for faith in a method that merely claims to approach equality of distribution of variables *in the long run*. How long the run must be, nobody can tell, and en route the randomization frequently produces inequalities that look impressive. Indeed, if we found in any small or moderate-sized samples exactly equal distributions of several variables that were independent of each other, we ought to suspect human interference.

Adjustment of Allocation after Randomization. If, after randomization, we detect an inequality of some feature between treatment groups, we may be tempted to make the groups more alike in this respect. If we are so tempted, we should remember the following five points:

1. If after randomization we examined a large number of independent features of the individuals in the treatment groups, we ought to expect the frequencies of some of them to differ greatly between the groups. If we call "rare" any difference that occurs in only 5 per cent of randomizations, we should expect about 5 per cent of these differences to be in the "rare" class.

2. In any particular experiment the purpose of the randomization is not to distribute individual variables (attributes or measurements), but the sampling units, each of which carries the resultant or net effect of a large number of different variables.

3. If we try to balance certain detectable features like sex or age and duration or severity of disease, it is not unlikely that we are throwing farther off balance some hidden features that may be far more important than those that we know about.

4. If we consider some features so influential that we wish the treatment groups to be homogeneous in that respect, we ought to use it as a basis for subdivision of our sample initially, and then randomize treatments within the subsamples.

5. If we have allowed any known feature (e.g., sex) to be randomized as part of the complex of variables, at the end of the experiment we can divide the data and compare treatment effects in the separate subclasses, e.g., males and females.

Q VI – 10. How are we going to insure that the randomization will be solely responsible for bias throughout the experiment?

To answer this fundamental question we seek in vain for detailed rules. After grasping some principles we must apply our imagination to the particular circumstances of our experiment, and we must exert constant vigilance.

The Blindfold Method. The best way of avoiding interference with the effects of randomization is of course the blindfold method – everyone who could in any way influence the results is kept in the dark regarding the treatment applied to particular subjects or specimens. Even in a blindfold drug trial there is no way, except education in scientific method, of preventing rank dishonesty, such as the illegitimate opening of the "emergency envelopes," i.e., the set of sealed envelopes, one for each patient, containing a statement of the compound that he is receiving, available in case he displays untoward symptoms and it is necessary to know whether he was receiving drug A or drug B. (He must, of course, be removed from the trial before the envelope is opened.)

Occasional leakages of information are bound to occur, for example when a pharmacist, having labeled a bottle with the patient's name forgets to remove the drug label before sending the bottle to the ward. Constant vigilance must be maintained by all participants.

Nonblindfold Experiments. Some clinical trials cannot be run blindfold because the secondary effects of the drug under test reveal the drug – for example, the moon-like swelling of the face resulting from the action of the corticosteroids. When the drug under test has to be injected there is naturally some hesitation to inject a placebo solution because of the possibility of infection, although with standard precautions the risk is very slight.

The Revealing Labels "A" and "B." It is surprising how many investigators think that they are conducting a double-blind trial when they use bottles of drugs labeled "A" and "B." If the observer sees, or thinks he sees, a difference in response between the A-treated and B-treated patients he is likely to be biased or to try to avoid bias. The difference in handling and evaluating the two groups may be subtle but very real. If the patients compare their experiences they, also, can produce biased responses. Moreover, if it is necessary in an emergency to find out what drug a patient is receiving, the drugs received by all patients in the trial will be revealed. Sometimes an attempt is made to mystify observers and patients by assigning two letters or label-colors to each drug; but this is still a treacherous "pseudo-double-blind" technique.

"Objective" Methods. If an experiment is not blindfold the rule is that objective methods of assessment must be used; but in fact it is very difficult to insure that a method is in all respects neutral to the treatment that a subject is known to be receiving. Functional tests, such as the measuring of grip strength in rheumatoid arthritis by asking the patient to squeeze a rubber bag connected to a mercury manometer, are readily influenced by the attitude of the examiner. In questioning a patient,

even if we use a set form of words, our tone or attitude can be easily affected by our knowledge; and there is much to be said in favor of asking the patient to write the answers to printed questions without giving him any explanation or help.

A blood pressure reading that we might accept and record for an A-treated subject we may doubt in a B-treated subject, and we may repeat the reading. Do we really know that our results would be the same if we took a second reading on all subjects?

A number of men, half of whom had received a treatment that was a possible preventive of nasopharyngeal inflammation, were lined up for inspection of the nose and throat by a medical officer who was skeptical of the treatment. After the inspection was over, an attendant remarked: "Did you notice, sir, that you spent much more time examining the treated cases than the controls?"

In a certain clinical trial the nurses were required to report small skin hemorrhages that might occur. The drug was to be administered intramuscularly into the buttocks. In the planning of the trial one of the clinicians was strongly in favor of a placebo injection in the controls because this would automatically insure that all patients had an equal opportunity of skin inspection by the nurses.

Destruction of Random Order. In laboratory work randomization can be ruined by a technician who is either improperly instructed or is unreliable. For example, in a certain experiment cage positions had been randomized and then it was found that the technician had, for his own convenience, grouped the cages according to treatment. In this instance the fault lay with the investigator, who always wanted the validity of his results to be "proved statistically," but thought that randomization was a statistician's useless fad.

Q VI – 11. If we are trying to find whether there is a difference in the effects of two or more treatments of any kind, what are the two types of error that we may make in drawing our conclusions?

On the one hand, we may conclude that the treatments differ in their effects when in fact they do not. On the other hand, we may conclude that there is, for our purposes at least, no difference in their effects when in fact there is a difference. This truism is the basis for a rather useful distinction between Type I errors and Type II errors.

Type I Errors. Let us suppose that we make it a rule to classify as "rare" the extreme differences – those that occur in not more than 5 per cent of randomizations such as card shufflings. After a treatment comparison, if the observed difference in the outcome is in the 5 per cent rarity class, we conclude that something more than the randomization was probably responsible, and in an experiment in which the randomization has been solely responsible for all the bias, the "something more" means the treatment difference. We can now visualize all the experiments in our lifetime in which, although we do not know it, there is no difference in treatment effects. If we adhere to the 5 per cent rule we

will commit a Type I error in not more than 5 per cent of these experiments.

More exactly, we ought to visualize the no-real-difference experiments conducted throughout the lifetime of an indefinitely large number of investigators, because any one of us may have bad luck in our randomizations – too many extreme differences. However, that need not worry us, because many of our experiments will be on treatments that do differ in their effects, and if we use the 5 per cent standard our total risk of erroneously inferring a difference will be much less than 5 per cent. If we use a 1 per cent standard of rarity, our risk will be less than 1 per cent.

In technical terms, the Type I error of a specified magnitude (e.g., 5 per cent or 1 per cent) is our risk of rejecting the null hypothesis when it is true; but it may be more useful to think of it as our risk of following a false clue.

Type II Errors. We must be rather careful in using this term. If an observed difference is not in our "rarity" class (5 per cent or 1 per cent or whatever we have chosen), we have no right to conclude that there is no real difference in treatment effects – we can never prove the null hypothesis true (Q VI – 1). We can merely say that the real difference, if it exists, has not produced a difference in our experiment that is big enough to satisfy our standard. Our verdict is "not proved," and in that verdict there is no error. However, after most experiments we act in some way or other as a result of our findings. If we act as if there is no real difference between treatment effects, when in fact there is a difference, we make a mistake, whether it has serious consequences or not.

We would like to know our risk of committing such mistakes, but the mere finding that a difference is not rare, on our 5 per cent or other standard, does not tell us our risk of making this other kind of mistake. To obtain a figure for this kind of error we need to ask a question such as: "If the real difference is such and such, and we take samples of size N, what is our risk of failing to detect a difference?" For instance, our knowledge of random sampling tells us what we will find if we have two large populations of subjects, A-treated and B-treated, like disks in two separate barrels, and if we then take from each population a strictly random sample of 20 subjects and repeat this process many times.

Let us suppose that the A-population contains 25 per cent X's (75 per cent not-X's) and that the B-population contains 50 per cent X's (50 per cent not-X's), that in each pair of samples of 20 we find the A-B difference in the numbers of X's and apply our 5 per cent standard of rarity. After repeating the process many times, we will find that we have judged the A-B difference "real" in only about 27 per cent of the sample-pairs. This is the percentage of our "successful" experiments, i.e., those in which we have detected something that exists in the populations. In the other 73 per cent of sample-pairs we have failed to find a difference that is large enough to meet our standard. We call 73 per cent our Type II error. It can be called our risk of accepting the null hypothesis when it is false, or our risk of failing to follow a true clue.

Lest we stray too far from the real world at this point, we should remind ourselves that when we are comparing A and B treatments, of any kind, on human or biologic material, we are not actually taking random samples from an A-treated and a B-treated population. As we proceed in a clinical trial, for example, the later patients may represent a somewhat different population from the one represented by the earlier patients with respect to their percentage of X's, even if they were all treated by A (or by B). All that we can hope to do is detect a kind of average difference, if one exists. However, the Type II error helps us in deciding what size of sample to take in order to have a good prospect of detecting a difference that we do not want to miss.

For instance, with samples of 100 and population percentages of X's 25 per cent and 50 per cent respectively (a difference of 25 percentage points) the Type II error is approximately 5 per cent. With populations containing 50 per cent and 75 per cent X's it is the same. As we pass farther from 50 per cent X's, but keep the sample size the same (or even somewhat smaller) and also the same population difference, the risk of error becomes less. Thus, with populations containing 5 per cent and 25 per cent X's and samples of 100 the Type II error is only 2 per cent. (The basis of these statements is revealed in Chapters XI and XII.)

The Power of an Experiment. In connection with the Type II error there is another rather useful term — "power." If our Type II error is 73 per cent, the percentage of "successful" experiments, 27 per cent, is called the "power" of that particular experimental procedure in the specified situation (composition of populations and sample size). If the Type II error is 5 per cent, the power is 95 per cent. The power is thus a numerical expression of what we have previously referred to as the "sensitivity" of the experiment.

It may be helpful here to mention a much-used symbolism, in which percentages are expressed as decimal fractions and Greek letters are employed. Thus, a 5 per cent Type I error becomes $\alpha = 0.05$. A 10 per cent Type II error becomes $\beta = 0.10$. With $\beta = 0.10$, the power is $1 - \beta = 0.90$.

Q VI – 12. If we intend to apply α test of "statistical significance" to our data, how will we know that it is α legitimate test?

A "legitimate" test is one that follows from the particular kind of randomization employed in the experiment. Having performed the actual randomization, we ought to know what series of randomization experiments would enable us to obtain the information that we need, about rarity of differences, and so on, without performing any arithmetical test at all. Then we ought to see how far the proposed test is equivalent to the series of randomization experiments, what assumptions underlie the test, and what is the evidence that these assumptions are safe with our data. These points are illustrated in later chapters; but here two general remarks may be made:

1. An investigator displays remarkable credulity, often unlike his other research behavior, if he performs (or requests, or demands) a statistical test without knowing its experimental equivalent, such as card shuffling.

2. Many research workers would be astonished if they knew the hidden assumptions in some of the tests and other mathematical performances that they take on trust. Sometimes they would see that the assumptions were inappropriate. More often, if they searched for evidence to justify the assumptions, they would find very little, either in their own data or in other data of similar type. They might then start to wonder what results they would obtain if half a dozen equally plausible assumptions were made, one after another, and used as a basis for the analysis of their data.

Q VI – 13. If a legitimate test is applied to our data and gives a verdict "statistically significant difference," what will it tell us?

Technical jargon, a useful shorthand, can do a great harm if its meaning, and especially its limitation of meaning, is not clearly understood. This seems to be the reason why the phrase "statistically significant" has done so much damage. The best way to elucidate jargon is to translate it. If a person says that a difference is "statistically significant," he means, or ought to mean, simply this: "Differences as large as this, in samples of this size, are so rarely produced by random processes alone, that I believe that this particular observed difference signifies (indicates, or points to) something in addition to a random process." This translation has purposely substituted "random processes" for "chance," because we are so often misled by colloquial meanings of "chance" – a mysterious force, or factors that we do not know about, or factors that we are unable (or too lazy) to correct for.

The 5 Per Cent Standard. It is for the investigator to decide what he will call "rare" in classifying random events. That matter is discussed under Q VI – 15. Here let us assume that he adopts the commonly used 5 per cent standard of rarity – the "5 per cent level of significance." To obtain an idea of what this means, let us consider the example given under Q II – 4. In 100 subjects (50 A's and 50 B's) there were, let us say, 35 "successes." If all 35 were in one group and none in the other, we would feel sure that randomization was extremely unlikely to be solely responsible. If 16 were in one group and 19 in the other, we would know that randomization would often cause such a trivial difference.

Somewhere between these two possibilities we must find a cut-off point such that all the more extreme differences must total to not more than 5 per cent of the total card shufflings. We could do this by actual card shufflings repeated, say, a thousand times; but we know so much about the effect of randomization in such simple cases that we can use the mathematical short cuts described in Chapter XII. Then we find that the required cut-off point lies between 13 and 12 successes in one sample (A or B) with the remainder (22 or 23 successes) in the other sample.

Hence, if the investigator's samples contained 12 or fewer of the 35 successes in one sample (and 23 or more in the other), he could say that the observed difference, if actually due to the radomization alone, would be in the "5 per cent rarity class of extreme differences." Translating this into jargon, he would say: "the difference is significant at the 5 per cent level."

The Probability P. To avoid the term "significant" the investigator could say "P is less than 0.05"; but this introduces another troublesome term, "probability." We are again misled by colloquial usage, especially the association of "probability" with the idea of prediction. Here, 0.05 is merely the decimal equivalent of 5 per cent, and we can define *probability* in this context as the relative frequency with which certain specified events (e.g., sample differences of a certain size or greater) occur in a series of randomizations, i.e., the frequency of these events expressed as a proportion of the total events produced by the randomization when it is repeated again and again. We are applying information that we possess regarding the effects of randomization, when it alone is acting, to the data from an experiment in which randomization was used, but to which something else (a treatment difference) has been added. We are not, as is a gambler, concerned with the use of the information for prediction.

We would perceive this more clearly if, whenever we saw "P less than 0.05" attached to an observed difference, we substituted a phrase like "Frequency in pure randomization experiments less than 0.05" or even "Randomization frequency less than 5 per cent."

Misconceptions Regarding "Significance" and P. When we conclude that a difference is "statistically significant" we should bear clearly in mind the following five points:

1. "Statistical significance" should never be used loosely, to mean proof of the reliability of the experiment or adequacy of sample size.

2. We must make our standard of "rarity" clear, either with each verdict or at the beginning of a report. Often "significant" and "not significant" are used without further specification to imply the 5 per cent level, while "highly significant" or "very significant" implies the 1 per cent level (P less than 0.01). This habit creates the impression that these purely conventional levels are almost laws of nature.

3. Statistical significance, even with a P value that contains many zeros after the decimal point, does not necessarily mean practical significance. There is often a reduction in human stature between morning and evening, but it is so slight in adults relative to their total length that for clinical purposes no allowance need be made for it. Some drugs, tested for pain-reducing properties, have produced statistically significant effects, but the effects have been too small to be of clinical value. We should always ask a question such as: "If there is a real difference in the effect of A and B, how large or how small may it be?" In the analysis of data, a significance test should not be the end of the road.

On the other hand, size alone does not indicate importance. When

someone looks at a set of figures (usually his own figures) and says: "The difference is not statistically significant, but it is clinically significant," it is difficult to know what he means. He may be relating what he sees to other information that seems to point in the same direction, in which case he is making a statistical judgment. Or he may be impressed solely by the size of the difference. Or he may be retreating to unverified clinical impression.

4. Statistical significance is not a mathematical proof that a real (population) difference exists. If we were living in a world in which differences between drug effects did not exist, our drug trials, if we conducted any, would still produce differences with $P = 0.05$ or less, and they would do so in 5 per cent of our trials. Because drugs in the real world often differ in their effects, when we meet a significant difference in a particular trial, we are justified in having some confidence that further study of the same kind of patient under the same conditions would reveal differences in the same direction. But the P value tells us nothing about how often this will happen or fail to happen. It would not do so, even if our original samples had been random samples of their respective treatment-populations, and if we sampled again from the same populations.

5. P values tell us nothing about the nature of the association between the variables that we are studying, or about the size of the differences in effect between the factors under test.

Q VI – 14. If a legitimate test is applied to our data and gives a verdict "nonsignificant difference," what will it tell us?

A translation of "nonsignificant difference" can be thus: "Differences as large as this, in samples of this size, are so often produced by random processes alone that these data do not convince me that anything more than a random process was responsible." "So often" means, of course, more often than the 5 per cent random frequency or other value that we have chosen as our standard. The verdict is simply: "Not proved." It is not a declaration that the factors under test have actually the same effect – that the two samples have come from the same population with respect to the dependent variables that we are studying. Nor is it a proof that the difference is "insignificant," i.e., of no consequence. It implies that, so far as we can tell from the data, a real difference, which we might find by taking larger samples, might be in the opposite direction from the difference found in our experiment.

Again, our experiment does not tell us much unless we ask a question such as: "If a real difference exists, how large may it be in either direction?"

Q VI – 15. If we are going to apply a test of statistical significance, what level of significance will we choose?

Translated, this means: "What is our definition of 'rarity' in randomization experiments? What is our cut-off point in a distribution of random frequencies which will determine our rejection of the null hypothesis?"

We must, of course, make our decision regarding the cut-off point

before we see, or even suspect, the results of our experiment. Otherwise, we may be biased, or our decision may be influenced by our fear of being biased. We may decide to accept the conventional cut-off point, $P = 0.05$ or, if we desire greater assurance, $P = 0.01$; but we ought always to know what our standard implies – what may be the consequences of our choice. Here we are helped by the concept of the Type I error, which is just another way of stating our standard of rarity. It specifies our risk of accepting as real the apparent evidence of a difference that does not in fact exist – our risk of rejecting the null hypothesis when that hypothesis is true. For example, it tells us how often we would accept one drug as more effective than another drug when we were comparing two drugs that did not in fact differ in their effects. In a series of laboratory investigations it tells us what proportion of the false clues we would follow.

We have to weigh this risk in relation to the particular circumstances of our study. There are no universal rules, but the following three points are worth noting:

1. It seems sensible to demand more convincing evidence (e.g., P less than 0.01 instead of 0.05) if the verdict of "significance" would run counter to previous experience or knowledge.
2. It seems sensible to accept less strong evidence (even P between 0.10 and 0.05) when the verdict would agree with other experience or knowledge.
3. The higher our standard of rarity – the lower the P value accepted as a cut-off point – the more likely we are to miss a real difference.

Even these simple guidelines are not easy to apply. Indeed, although the scheme of error risks (Types I and II) is a useful concept, it does not provide us with a definite numerical standard in most biologic or medical research, as it does in the sampling of industrial or agricultural products. In those areas the producer can find in dollars the cost of the two types of mistake that he may make in estimating, from a sample, the percentage of defective items (e.g., lamp bulbs that burn out too soon) in the whole batch – (1) underestimation, which will lead the consumer to seek a more reliable product; (2) overestimation, which will necessitate discarding the whole batch or selling it at a lower price.

Counting the cost in medicine, even in drug testing, is much more difficult. Even if an investigator says that he has set his risk of a Type I error (his significance level) at a certain value, the best way to discover his real standard is to see what he does after he has obtained his results. Having found a certain difference in outcome between two treatment groups, does he act as if it were due to the treatment difference, or as if it were fortuitous, or as if he really did not know?

Q VI – 16. If the verdict of a test is "significant" at our predetermined level, what will we do in consequence thereof?

This question, considered before we start the experiment, should prompt us to think again of the immediate purpose of our research – the direct

question that it is designed to answer – and to ask ourselves whether a verdict of "significant" under the circumstances of our experiment would give us an adequate answer. No specific guidance can be given, but it is helpful to reflect on six topics: (1) rare events, (2) credibility of interpretations, (3) multiplicity of causes, (4) associated agents, (5) group effects, and (6) experimental proof. Although some of the examples are taken from surveys or even from casual observations, and not from experiments in the strict sense, the message that they convey is equally important in experiments.

Rare events. Very unusual events can occur by chance, as is illustrated by the following three quotations:

> Boswell in his *Life of Samuel Johnson* tells us that an acquaintance of Johnson "observed, as something remarkable that had happened to him, that he had chanced to see both No. 1 and No. 1000 of the hackney-coaches, the first and the last." "Why, Sir," said Johnson, "there is an equal chance for one's seeing those two numbers as any other two."
>
> "The chance that a pack of cards shall be dealt in *any* assigned order is the same – less than one in a billion; we suspect a joke or a fraud only if the particular order in which the cards fall suggests design" (Greenwood, 1944).
>
> "The 'one chance in a million' will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*" (Fisher, *The Design of Experiments*, sect. 7).

These statements do not imply that, having met a rare event, we should never seek for causes other than chance; but they show that, however long we may seek, we may not find such causes, because they may not exist. It is well to recall this when we read letters in medical journals recording an unusually high incidence, in the same district at the same time, of a reputedly noninfectious disease such as appendicitis, or reporting the simultaneous occurrence of cancer in two or more unrelated persons in the same household.

Credibility of Interpretations. However much evidence there may appear to be in favor of a certain causal interpretation, we may find ourselves unable to accept the interpretation. As Greenwood (1944) pointed out, if we stuck postage stamps to the beds of some patients (without knowledge of patients or staff) and found that all those patients recovered, whereas all those without stamps died, we would nevertheless not attribute the recovery to the postage stamps, because the relationship would not be credible, or conceivable, or rationally acceptable.

The criterion of credibility can, however, sometimes mislead. For example, when Jenner used cowpox vaccine to protect people from smallpox, his evidence was rejected by some of his contemporaries because it was inconceivable that one disease should protect against a different disease (Greenwood, 1944). The rationality of the method was discovered much later when the viruses of the two diseases were shown to be related.

Multiplicity of Causes. In an abstract form we can describe a cause-

and-effect relationship by an expression such as: "If A, then X; if not A, then not X." Actual phenomena, however, are seldom as simple as this. When a person becomes infected by an organism we think of his illness as due to that organism; but it is due also to his lack of resistance, and this is due partly to environmental factors, including nutrition and numerous events affecting his health in the past.

In pure research we may try to discover all causal factors, but in applied research we commonly choose as the cause of a phenomenon one or two major factors, often those with which we can most easily interfere, in order to prevent or cure disease. There is danger in this, however, because there may be other equally important factors. For example, the spectacular success of bacteriology led us to concentrate on the "seeds" of disease, bacteria, and pay too little attention to the "soil," the patient's constitution, physical and psychologic.

Associated Agents. There is a well known story of a man who got drunk first on scotch and soda, then on brandy and soda, then on rye and soda, and then blamed the soda, ignoring the *associated agent* or *concomitant factor*, alcohol. This fallacy is common in medicine for two reasons:

1. Our ignorance. For example, certain of the benefits of cod liver oil were long attributed to the oil itself, whereas we now know that they are due to vitamins in the oil.

2. The complexity of phenomena. Most clinical treatments comprise many elements, which a clinician cannot study one by one, as would a laboratory worker. A physician can sometimes omit certain elements from his treatment and watch the effect of the rest, but the surgeon can hardly perform a series of "dummy" or partial operations from which various parts of the full operation are omitted.

For our immediate purposes, a wrong explanation of a successful treatment may not matter, but in the long run it may be very misleading. The important thing for an experimenter in any field to remember is that "Treatment A" means the composite of items so labeled.

The associated agent that most commonly misleads the clinician is *the patient's mind.* The psychologic element accounts in many instances for the following phenomena:

1. The success of a certain treatment with some patients although it fails with others. Many patients try to please, or at least to avoid discouraging, their physicians. Others resist his efforts, sometimes because it pays them, either psychologically or financially, not to get better.

2. The failure of a treatment in a certain patient after its initial success.

3. The success of a treatment in the hands of one physician, although it fails in the hands of another.

4. The success of some new and highly advertised remedies.

The risk of our being misled by the psychologic component of therapy is of course the chief reason for the use of placebos in testing drugs. When they are impossible we have to depend on two less reliable methods: (a) "objective" observations, often questionable because we seldom know the extent to which the mind can affect the body; and (b) long-term follow-up, which is useful because a psychologic effect is likely to wear off sooner than a physical effect.

Group Effects. Although an intergroup difference in outcome may be clearly attributable to the difference in treatment applied to the two groups, we must always remember that we are talking about groups. For example, a patient who had improved on the more effective drug A might have improved even more on the less effective drug B; and this is equally true if we have used each drug in succession on all patients in a cross-over design.

Experimental Proof. "In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result."

If for the phrase "statistically significant" we substitute a word like "consistent" or "similar" this statement is remarkably like the assertions of some experimenters who distrust "statistical" methods and who repeat their experiments until they are convinced that they know how to produce a particular effect. It may, therefore, be a surprise to them, and also to some devotees of significance testing, to know that the quotation is from *The Design of Experiments* by the statistician who did most to develop and disseminate significance tests, R. A. Fisher (later Sir Ronald Fisher). To those who were directly acquainted with Fisher's attitude to these tests and his close contact with experimenters' problems, the quotation contains no surprise. It would, I think, be fair to express his concept of the function of significance tests as follows. They help us because they enable us to separate two classes of results: (a) those which, from the evidence of the particular investigation, could well be fortuitous; (b) those which would be hard to explain by random processes alone.

Q VI – 17. If the verdict of a test is "not significant" at our predetermined level, what will we do in consequence thereof?

Although we recognize that a verdict of nonsignificance simply means "not proved according to our standard of evidence (our chosen significance level)" we may act as if "not significantly different from" is synonymous with "the same as."

In a clinical trial two drugs may not have differed "significantly" in the features that we have studied. Perhaps we cannot wait for further evidence before using one of the drugs, or perhaps we have concluded that the difference, even if it exists, is probably not great enough to be important. In such cases our choice of drug will sometimes be determined

by other considerations. For instance, if only one of the drugs produced undesirable effects, we would prefer the other one. Cost, difficulty of administration, our greater familiarity with the standard drug than with the new one, other people's experience with the new drug – these and other factors may influence our choice. If there were no objection to either drug, we would be likely to choose the one which, in the trial, seemed to be more effective.

An example of another kind of behavior occurred in an investigation of the treatment of traumatic shock during World War II. Dogs were experimentally injured by a certain technique and then 15 of them were kept at an environmental temperature of 95 degrees Fahrenheit, whereas 10 were "cooled to an equivalent degree." Of the warmed animals 11 (73 per cent) died; of the cooled animals only 4 (40 per cent) died. Pure random assignment of the 15 deaths among 15 animals labeled "A" and 10 animals labeled "B" would produce in about 20 per cent of the randomization experiments differences equal to and greater than those found in the actual experiment. In other words, the difference was far from significant at the 5 per cent level, because P was about 0.2. That is, if we accept the observed difference as indicative of something more than a random process, we are adopting a standard that would lead us to follow something like 20 per cent of false clues.

The investigators actually took that risk, although apparently unaware of it because the report did not mention a significance test. With a larger number of dogs they obtained clear evidence of the benefit of cooling. Therefore, disregard of the risk in this case was justified. Disregard of such risks, i.e., the acceptance of a lower standard of significance than usual, is always legitimate provided that the reason is well defined. The actual reason here is unknown, but it might be one or more of the following:

1. The possession of some other knowledge of the physiologic effects of cooling that would agree with, or explain, the observed difference.
2. The ease with which a larger experiment could be done.
3. The fact that if cooling did reduce mortality, it might be very important in treating human shock.
4. The large size of the difference, $73 - 40 = 33$ per cent. This apparently impressed the investigators, because when they first announced it they did not even mention the sample sizes. Such frequency differences are often impressive to laboratory workers who are used to small percentage errors in measurement data.

Q VI – 18. What will make us decide to terminate the experiment?

Fixed Sample Sizes. In the experiments discussed so far, the sample sizes are fixed in advance. When we have reached these numbers of subjects and apply an appropriate test to our results we know our Type I error. During the experiment, however, even if there is no real difference in the effect of the treatments under test, we nearly always observe fluctuations in the data. Sometimes one treatment seems to be favored

and sometimes the other, just as if we had cards, some marked "S" and some marked "F," in a well shuffled set and turned them up one by one.

Let us suppose that as the experiment proceeds we get the impression that a treatment difference is emerging, that we then apply a significance test and find that P is less than 0.05. That result has no meaning with reference to our data. We have analyzed the samples, accumulated up to that point, as if they were strictly random samples, whereas we have in fact chosen them because of their contents. If we do this we will obtain far too many "statistically significant" differences in our experiments.

Indeed, we can obtain as many "significant" differences as we wish if we are willing to go on trying. This can be shown by using cards or random numbers as "subjects" of the "experiments," or by an arithmetical lay-out that represents the behavior of a random process. In these experiments there is no difference between the "treatments"; they are merely labels such as A and B. We make it a rule to stop the "experiment" whenever we have reached a verdict of "significant" at any predetermined level (e.g., 0.05), but to continue to use more "subjects" (cards or numbers) whenever we have not reached that level.

By this procedure we could reach a verdict of "significant" in all our experiments. Some of the experiments would be exceedingly long; but the spurious verdicts can start whenever a laboratory worker uses the "try, try, try again" method. Having found that the first batch of animals does not give a statistically significant difference, he adds a few more animals. If the verdict is still "not significant" he adds still more animals, and tests again.

Perhaps the damage done by this misuse of significance tests is not tremendous, because the investigator often gets tired after three or four tries. His true risk of Type I error may be about 10 per cent, instead of less than 5 per cent, as he thinks it is. The point is that his actual risk of error depends upon his impulse or whim in each individual experiment. This is one of the many ways in which the use of statistical tests in laboratory and clinical research makes the vaunted claims of precision for statistical techniques rather ludicrous; and the blame, I think, lies largely with those of us who have produced books of elementary statistics.

There is, of course, another danger in stopping an experiment because of the results that are emerging. If we have an impression that no difference between treatment groups is appearing and then we apply a test, we run the risk of having too many "nonsignificant" differences, when in fact a real difference in treatment effect is present.

These warnings do not imply that we must deprive ourselves of the right to stop an experiment whenever we wish, in view of what appears to be coming out of it, especially in relation to other knowledge that we possess, or because we see disadvantages in continuing it – for example, an excessive number of undesirable effects in a drug trial. It simply means that if we intend to play the "significance testing game" we must

obey the rules. The sample size must be determined by something unrelated to the data that it is producing, unless we use the "sequential design" discussed below.

Samples Determined by Amount of Material and Time. In order to perform a fixed-sample-size experiment, suitable to give us an estimate of error risk at the end, we do not need to specify in advance the actual sample size. We can decide to put into the experiment all suitable subjects available during a certain period of time. The danger in this plan is, however, that the intake of subjects may be speeded up or slowed down as a result of the figures that emerge during the experiment. For example, a physician who sees no exciting difference between the effects of two drugs in a clinical trial may relax his efforts to secure suitable patients.

The Sequential Design. This method is much more akin to the experimenter's centuries-old step-by-step procedure than is the fixed-sample-size method, but it has been only recently developed as a systematic design, and it is still undergoing changes. When it was devised during World War II, it was for a time an official secret, because in the testing of materials and processes it enabled verdicts to be obtained from smaller samples, and therefore more quickly, than did fixed-sample-size methods. In many drug evaluations, a decision based on the fewest possible patients is desirable for ethical as well as economic reasons.

The distinction between the sequential design and the traditional step-by-step method of experimenters is that the sequential method enables us to set in advance our Type I and Type II errors, whereas the old method does not allow us to estimate our error risks either before or after the experiment. The sequential design does not graft the fixed-sample-size significance tests illegitimately on to the step-by-step method, but provides decision rules of its own. These rules are developed by complicated mathematical techniques, but that need not disturb us, provided that we know three things:

1. What the method tells us in terms of the random sampling of disks from a barrel.
2. The difference between our experimental circumstances and the disks in the barrel.
3. The limitations of the sequential method in medical research.

In Chapter XVI (sect. 3) is an example of the rules for a sequential experiment, and their implications in terms of random sampling are discussed. Here we look at the general procedure, as exemplified by a drug trial.

Sequential Drug Trials. We have randomly assigned patients to treatments A and B. When the result from the first patient comes in we hold it until we can pair it with the first result from a patient on the other drug. We then compare these two with respect to the measurement of drug effect that we have decided upon, the change in a certain variable since the pretreatment examination. If both patients show the same

amount and direction of change they tell us nothing about the difference in drug effect if one exists, and we discard that pair. If we assess the patients broadly as "better," "worse" or "no detectable change," we may have many such "tied pairs." (The pairing does not imply initial matching, but this can be done if there is a good reason for it, and if it is feasible, e.g., if in an animal experiment litter mates are available.)

Each "untied" pair shows evidence that it is apparently in favor of A or of B. If the first pair shows an A-preference we can count it as a score of +1. If the second untied pair, obtained in the same way, also shows an A-preference, the score becomes +2; but if it shows a B-preference the net score is zero. We proceed in this way, and for each number of untied pairs the rules tell us whether (1) we have accumulated a large enough (plus or minus) score to say that A is significantly better (or worse) than B, or (2) the score is so small that we can pronounce a verdict of "no significant difference," or (3) we must continue the experiment.

Some Drawbacks of the Sequential Design. Before deciding to use a sequential design we must find out how much we are likely to save, in sample size, by doing so (see Chapter XVI); but even a great potential saving may not compensate for the disadvantages of the method. The principal drawbacks can be expressed in terms of a clinical trial, but translation into terms of any other kind of experiment will be obvious.

1. The original type of sequential design is "open-ended," i.e., although it usually requires smaller samples than the fixed-sample-size design, the experiment may continue indefinitely without reaching a verdict of either "significant" or "not significant." Modifications, called "restricted" or "closed" designs, have therefore been devised. Their minimal sample-size requirements are not as low as those of the open-ended design, but in using them we always know at the beginning the maximum sample size that may be necessary. Such knowledge is often very desirable in drug evaluation.

2. The sequential design is of no use unless the period of treatment of the individual subject is very short compared with the total length of the experiment. In a clinical trial if the treatment lasts only 2 or 3 days we can withhold patients until we have obtained definite results from the first dozen or more; but if the therapy has to last 6 months all available patients may have been admitted before the first few have completed their period of treatment.

3. A sequential experiment is intended to answer the specific question for which it is set up. If in addition we desire answers to some questions about other variables, and if those variables are closely related to the variable on which the sequential experiment was based, we will obtain too high "significance" estimates. We have decided to stop the experiment because of the result shown by the primary variable. Therefore, if we now treat the samples as in a fixed-sample-size experiment for comparison of other closely related variables, we commit the error already discussed in the section on Fixed Sample Sizes.

4. Whether sample size is predetermined or determined sequentially, a small sample contains less information than a large sample. Regarding the sequential design, a clinical investigator has written: "It seems to me unlikely to supplant completely the large-scale therapeutic trial. . . . When a new treatment is introduced for an important disease, assuming the treatment is not such a radical advance as to make formal testing unnecessary, there are several advantages in making a big study. First, . . . a big study permits of analysis to show the scope of the treatment within the disease. Secondly, a big study is often an advantage in permitting one to take stock of secondary aspects, such as complications of the new therapy, which may be of great importance" (Truelove in *Medical Surveys and Clinical Trials,* edited by Witts).

Q VI – 19. Are we going to make more than one comparison in the same data? If so, are significance tests legitimate?

We can look at two kinds of multiple comparisons: (1) Comparisons suggested by the data, (2) Preplanned comparisons.

Comparisons Suggested by the Data. Let us suppose that we had a set of cards representing treated patients, some cards marked "S" (Success) and others "F" (Failure), that we shuffled them and then marked half of them "male" and the other half "female" without regard to their S and F marks, then shuffled again and marked half of them "young" and half of them "old," then shuffled again and marked half of them "treated early" and half of them "treated late," then reshuffled them and marked half of them "sedentary work" and the other half "heavy work," then reshuffled them and marked half of them "brown eyes" and the other half "blue eyes," and so on, through many contrasts.

If we recorded the results after each shuffling we would sometimes find that many of the F's had been thrown into one class and few into the other class. If we applied a "significance" test in those instances we would find in some, or perhaps in all of them, a verdict of "significant" at the 5 per cent level, or even at a higher level. Indeed, we ought to expect 5 per cent of such shufflings to give such a verdict, because that is what "significant at the 5 per cent level" really means.

And so, if we make comparisons in any set of real data we can always expect to find some that are significant, even if there is no real association between the dependent and independent variables. If we pick out for testing only those differences that appear striking, we may find every time a "significant" difference. The real value of searching for big contrasts lies in the discovery of hypotheses to be tested by a new investigation. When we have the results of that investigation we must, of course, present them separately from the data that gave us the hint. If we pooled the new data with the first data, and tested the combined data, our test would be spurious. The samples would not be random but selected because of part of their contents, the part contributed by the first data.

Preplanned Comparisons. In a trial of streptomycin plus bed-rest against bed-rest alone, pulmonary tuberculosis patients were not divided into groups according to temperature (as an index of activity) at the

beginning of the trial (Medical Research Council, 1948); but from the outset the investigators planned to divide the data into several initial-temperature groups in order to compare patients' progress in the different groups and also to see whether the streptomycin versus bed-rest contrast differed according to the initial activity of the disease. In such comparisons, if the relationships of the dependent variables (e.g., change in the patients' condition) and the independent variables (e.g., initial temperature) were purely fortuitous, no more than the usual 5 per cent of the comparisons would in the long run yield P values of 0.05 or less.

In clinical trials several measures of effect are commonly used – for example, in rheumatoid arthritis, strength of hand grip, number of painful joints, time required to walk 50 feet, and other measures. It is very unsafe to assume that these provide independent evidence of the effect of a drug on the phenomena that cause the symptoms or disabilities. Indeed, if there were perfect correlation among the measures of change, e.g., if an increase in grip strength of 40 mm. of mercury always accompanied a reduction in the number of painful joints by 10 and a reduction of 2 seconds in the walking time, a "significant" difference in any one of these measures would always be accompanied by a "significant" difference in the others. And yet all of these differences between the treatment groups might be due to the one, rather uncommon, random allocation of patients to therapies.

Although perfect correlation between the measures of change does not exist, some degree of relationship is commonly present, and allowance can, in part at least, be made for it. For example, we can ask: "If there were no drug-placebo difference in the change of grip strength, does it appear that there would be a drug-placebo difference in the change in numbers of painful joints?" Such questions are seldom of great concern in a drug trial; but in other experiments they may lead to important discoveries.

Q VI – 20. If we are comparing more than two treatments, how may significance tests mislead us?

It often seems desirable to compare, in one experiment, three or more treatments, each on a separate group of subjects. If we compare A with B, A with C, and B with C, with 20 animals in each treatment-group, it would seem that we can obtain three answers from 60 animals, whereas if we made the comparisons in three separate experiments we would require 120 animals. However, the matter is not so simple.

Comparison of Extremes. First let us consider an experiment in which there is no natural grouping of the treatments, such as there would be if A and B were drugs, with C a placebo, or if A and B were chemicals of similar composition, with C differing in some specific way from both of them. Let us suppose that the results (e.g., percentage frequencies or average values) show the largest difference between A and B, with C intermediate in value. If we select the A-B difference, apply a two-sample test and pronounce the difference "significant" because P is less than 0.05, our verdict will be spurious. We are not comparing random

samples but samples selected because of their contents. Our action is something like picking the winner *after* seeing a horse race. No one would accept our bet, but we may win acceptance for our significance-test verdict if we submit our report to the right journal.

The first thing to do in such a multiple-treatment comparison is to look at the intergroup variation as a whole and ask: "How often does a randomization like the one used in this experiment produce as large intergroup differences as this?" There are tests that help us to answer this question, and they are mentioned in Chapter XII (frequency data) and Chapter XIV (measurement data).

If we decide that the intergroup variation is not "significant" according to our standard, this may be all that we wish to know; or we may wish to pursue an interesting apparent difference in the data and devise another experiment to test its genuineness. If, however, we find that the intergroup variation meets our standard of "significance," we naturally ask: "Which treatments differ from which other treatments in their effects?" This kind of question is not easy to answer. Professional statisticians have proposed several different answers – different methods of analysis – and have, of course, argued with each other about them. Therefore, the rest of us should be chary of using any of the proposed methods. Each method may be correct in the circumstances, and with the assumptions, postulated by its proponent, but in biologic and medical research we often know little of the circumstances beyond what we see in front of us, and we often do not know what assumptions are valid.

If we have decided that the intergroup variation is "significant" and if we have to act on the result of the one experiment (e.g., to choose one therapy out of three or more), our "best bet," although we cannot attach a "probability" to it, is to assume that the most extreme difference represents a real difference. Here we are acting on the principle that a sample is more likely to be near to the center of its population than to be far off the center. We can trust this principle best if our samples are large, because a big difference between small samples can be less real than a smaller difference between larger samples.

All Possible Comparisons. The suggestion that, for purposes of action, attention be confined to the extreme differences hardly satisfies someone who has put three or more treatments into his experiment, expecting to obtain a definite statement about each treatment. Having learned some simple two-sample tests, he may proceed to test each treatment against every other treatment in turn. Perhaps we can see the danger in this method if we imagine that someone takes a strictly random sample of the statures of the men in a certain city, and then compares it with random samples of men's statures from half a dozen other cities, one sample per city.

Let us suppose that the populations of all seven cities are actually identical with respect to men's statures, but that the investigator obtained from his first city a rather rare type of random sample, i.e., one whose average stature differs greatly from the true average of that city. He

may then find "significant" differences between this sample's average and the averages in the samples from each of the other cities. He will, at least find P values that are smaller than what he would find if he had obeyed the rules that make the significance test valid, i.e., if he had taken a fresh random sample from the first city every time that he compared it with another city. Clearly, the same kind of thing can happen in a multiple treatment-comparison.

Multiple Intergroup Differences. In a comparison of three or more treatments (A, B, C . . .) a verdict of "significant intergroup variation" does not tell us a great deal. It justifies our belief that, if we could find (or build up) the corresponding treatment populations, we would discover that two or more of them actually differed with respect to the variable that we are studying; but this could be true in a number of different ways. The A, B and C populations might all differ from each other, or A and B might be identical but differ from C, or B and C might be identical but differ from A. In any of these situations the variety of combinations of possible population values might be infinite, so far as we can tell from our actual samples. Usually, therefore, if we wish to define more clearly the individual treatment differences after a multiple comparison of the type discussed here, we must perform more experiments.

We can proceed a little farther in the kind of experiment in which treatments A and B are alike in a certain respect and differ in that respect from treatment C, e.g., two drugs versus a placebo. In the analysis we start with a comparison of A and B. If they do not differ significantly we can pool their data and compare with the data from C – provided that we realize clearly what we are doing. In effect, we are saying: "We are not convinced that there is a difference between the effects of A and B. If A and B were actually equal in their effects, would this experiment indicate that C differed from them?"

It is important to realize that once we have used the information from our experiment in the analysis just described, if we use it over again to make some more comparisons (e.g., A versus C, or B versus C) we run into the same danger as in the imaginary stature-survey of the seven cities.

Let us suppose now that a test has shown that A and B differ significantly (A greater than B) and that the value (e.g., average) for C is lower than that for B. It seems appropriate to compare B with C and accept a verdict of "significant." If, however, the verdict is "not significant," we should realize that this verdict and the "A-B significant" verdict may have been due to the same cause. The effects of A and B may not actually differ, and both may differ from C, but the randomization may have, as it were, pushed the B-sample so far below the A-sample that it came near to the C-sample.

All these possible interpretations after multiple treatment-comparisons, of the kinds discussed here, are rather frustrating. The method does not carry one as far as might be expected. Factorial designs (Q IV – 8) carry us much farther.

Q VI – 21. How will we estimate what our results really tell us, numerically, about the population represented by our sample?

This question recalls the quandary discussed under Q III – 9. On the one hand, unless our sample of animals, patients or other sampling units is a strictly random sample of its population, we cannot derive from it a population-value estimate that has known reliability, i.e., an estimate with a known risk of error. On the other hand, in most medical research we cannot obtain strictly random samples of the populations to which we desire to apply our results; or, if we do possess samples that are close enough to random samples for our purpose, we do not know that they are. And yet we badly need to know how far we may be misled if we accept our sample estimates at their face value.

The Problem of Nonrandom Samples. In the Medical Research Council's (1948) 6-month clinical trial, comparing the treatment of pulmonary tuberculosis by bed-rest alone with the treatment by bed-rest plus streptomycin, there was a difference in outcome (statistically significant at the 5 per cent level) in the mortality (case fatality rates): 14 of the 52 purely bed-rest patients died, but of the 55 patients who received streptomycin only 4 died. No one would deny the clinical importance of this difference, but the next question should be: "What do the 4 deaths in 55 streptomycin-treated patients (7.3 per cent) tell us about the case fatality rate that would be found in a large population represented by these streptomycin-treated patients?" More precisely, we should say: "What case fatality rate might be approached if more and more of this kind of patient were treated in the same way?"

The basic features of the population are given by the description of the sample: patients of both sexes between the ages of 15 and 30 years inclusive, with bilateral pulmonary tuberculosis, bacteriologically proved and unsuitable for treatment by collapse of the lung (e.g., by introducing air into the pleural cavity). The definition was further restricted by the fact that the trial was conducted at seven specified hospitals in the United Kingdom during the year 1947. We would make the definition much more specific by adding many of the details recorded in the report, such as temperature on admission; but we would still have no absolute assurance that the 55 observed streptomycin-treated patients would be equivalent, for the estimation of possible error, to a strictly random sample of a population so defined. Indeed there was no such population, because during the period of the trial it was stipulated that all eligible patients at the seven centers be enrolled in it.

The 1948 and 1949 samples from the same centers and defined in the same way might indeed differ from the 1947 sample by more than random differences, i.e., they might represent different populations; but if such a possibility prohibited any sort of estimate from the 1947 sample it should by the same reasoning have prohibited the clinical trial itself, and it would prohibit most other clinical trials and many other kinds of experiments. The streptomycin investigators clearly refused to be blocked in this way, because they knew that the disease and other circumstances

in chronic pulmonary tuberculosis do not change so greatly from one year to the next as to stultify their 1947 experiment.

Random Processes in Sample Formation. The more specifically we describe our sample, by race, sex, age, condition at start of therapy, strain of organism in certain diseases, and by other major factors that influence a patient's progress, the nearer we come to a situation in which the differences in outcome between successive samples, so defined and treated alike, are due to random processes, i.e., the action of a number of factors that are independent of each other. Some of the patients in the 55 who received streptomycin would be, from the start, likely to do well and others unlikely to do well, for reasons not yet discoverable by any method of investigation; but the presence of each patient in the sample of 55 would be determined by a combination of circumstances stretching far back in his own and his ancestors' history, a different combination for each patient. This reminds us of the mixing of disks in a barrel during a random sampling experiment.

We next visualize the patients in the hospital during the trial, subject to many forces, independent of the state of their tuberculosis and independent of each other. One patient cannot "stomach" the hospital food. Another reacts in the same way to a certain nurse. A third patient finds both delectable. We add to this picture a virus that produces an attack of influenza in one patient but leaves another unscathed, regardless of their attitudes to food or nurses. Then we add a physician who is very perceptive and skillful in attending to even minor elements in patient care, and another physician who is less so. We could continue to add factors indefinitely, each of which could contribute a little to help or hinder a patient's response to the streptomycin. Some patients would receive mostly helpful factors, others mostly hindering factors; many would receive mixtures of both kinds. Again we are picturing a more or less random process.

The streptomycin investigators could therefore have said: "Unless we trust, provisionally, the uniformity of Nature we can get nowhere in drug evaluation. Let us assume that there were no major differences between our group of 55 patients and other groups at other times and in other places. Even so, there would be differences due to random processes, because random processes are part of the uniformity of Nature. Let us find out how much or how little our 7.3 per cent case fatality rate in 55 patients would tell us if random processes were the only cause of intersample differences."

Systematic Differences. In this imaginary quotation the phrase "no major differences" is too indefinite to be synonymous with "purely random differences." The phrase "no systematic differences" is preferable. "Systematic" is used here in the same sense as when we speak of the systematic difference between two instruments designed to measure the same thing. We estimate its size by paired readings on the same material, one reading by each instrument; but the reading difference is not necessarily the same in all the pairs, because fluctuations of the instruments and observer

can occur. We increase the precision of our estimate by taking the average difference from more and more pairs.

A systematic difference that would nullify the streptomycin data of 1947 would be the presence in subsequent years of patients whose tubercle bacilli were resistant to streptomycin, a phenomenon that has occurred during the use of many antibiotics. To translate the "systematic difference" into the terms of a disk-sampling experiment, when we first start sampling we can imagine that 10 per cent of the disks in our barrel are marked "D" (death) and 90 per cent "S" (survival). After a while somebody, unknown to us, slips into the barrel a batch of disks of which 40 per cent are marked "D." Our subsequent samples, although still strictly random, will be affected to a greater or less extent by the systematic change in the population.

Estimates of Population Values. The disk-sampling experiment would provide the streptomycin investigators with the most useful form of question regarding the reliability of their sample: "If our sample of 55 streptomycin-treated patients, containing 4 deaths (7.3 per cent) were a strictly random sample of its population, what would it tell us about the percentage mortality in the population."

We could find the answer to this question experimentally. For example, we could put in a barrel several thousand disks, of which 20 per cent were marked "D," and take a thousand random samples of 55 disks. If we found that samples of 55 containing 4 D's or fewer were "rare" by our definition of the term, e.g., if they comprised fewer than 5 per cent (or 2.5 or 1 per cent) of the total samples, we would conclude that the mortality in a population randomly represented by the streptomycin-treated sample was unlikely to be as high as 20 per cent. Then we could do another sampling experiment on a population containing 15 per cent D's. If we found that 4 deaths in 55 was not in the "rare" class of sample from that population, we would accept 15 per cent mortality as a possible value.

Then we could take a number of other populations, working up from 15 per cent and down from 20 per cent, until we discovered a boundary or limit between the population percentages that we would reject as unlikely and those that we would accept as possible. Chapter XI shows that we do not need to perform the experiments with actual disks, because we can do them on paper by using the binomial expansion, and that even that labor is seldom necessary, because tables such as our Table I give direct answers about limits of population values. For example, if we use 2.5 per cent as our standard of rarity in random sampling, the answer to the streptomycin investigators' question is: The percentage mortality in a population randomly represented by a sample of 55 with 4 deaths (7.3 per cent) might lie anywhere between 2 per cent and 17 per cent. (The estimation of population values from samples of measurement data is discussed in Chapter XIII.)

Such estimates, made after any observation, are minimum estimates of our present ignorance because they allow only for random sampling

variation. If, when we use our present estimate later on, there is also a systematic difference from the present conditions, the new population value may be far outside the limits that we have now estimated. But even our present minimal estimate of error is often very valuable. It may show that even at its best under present conditions a certain therapy would not be very effective; or that even at its worst it would be better than the therapy that we have hitherto employed. Often the chief value of the estimate is that it reveals how little our sample has told us.

Q VI – 22. Is a "null hypothesis" an appropriate concept for our experiment?

After a long chapter with the null hypothesis as its principal text, this seems to be a rather belated question. It has been delayed because it is necessary to see the implications of the null hypothesis before considering its applicability in particular cases. We can look at two types of cases: treatment comparisons and the study of observer and instrument differences.

Treatment Comparisons. It must be admitted that we seldom start an experiment with a completely open mind as to whether a particular hypothesis is or is not true. We often start a drug trial because the drug (or perhaps a relative of it) has apparently shown "promising" results in a few patients or animals. Our original hypothesis, therefore, might be that the difference in percentage of S's in an A-treated population would exceed the percentage of S's in a B-treated population by at least 20 percentage points. We could, however, test this by setting up a null hypothesis that the true difference was just below 20, say 19 percentage points. We could demonstrate, by a method based on the random sampling of disks from a barrel, how often various sample differences would occur if the 19 per cent hypothesis were true. We could classify as "rare" the largest differences (e.g., the upper 5 per cent), and then we would find out whether the A-B difference in our drug experiment was so large that it would be in the "rare" class. If it were so large we would reject the null hypothesis.

There are many such situations in which we can use our past experience, intuitions or convictions in the framing of a suitable null hypothesis; but we naturally ask: "Is there not some way in which we can express quantitatively our belief, based on previous experience, that a certain hypothesis is true (or false) and then combine it with a probability value that we derive from our experiment?"

Statisticians also have asked such questions, but the results so far have been chiefly rival theories. They are attempts to set inductive inference – the process by which we develop new knowledge – on a rational foundation, and to make it more efficient. Mathematical models are developed, containing such quantities as "prior probabilities" and "likelihoods." They necessarily involve assumptions and simplifications – exclusion of many of the complexities of the real world. To this in itself there can be no objection, because all mathematical models, such as $2 + 2 = 4$, are abstractions. The great distinction between the simple-

addition model and the models proposed for inductive inference seems to be that we have learned by experience where the simple-addition model is applicable, what it does for us in the real world. Until proponents of theories of inductive inference ("statistical inference") can propose also some way of properly testing the theories in the real world, an investigator need not feel guilty if he disregards them, unless he is called upon to help in testing them.

Observer and Instrument Differences. Some workers who are, very rightly, concerned about problems of instrument differences and observer differences make careful studies of these. For example, two observers may make repeated blood pressure readings on a number of subjects, each observer using the same two sphygmomanometers. Unfortunately, however, they sometimes analyze their data by "significance" tests, i.e., they set up null hypotheses (no interobserver difference, no difference between instruments) and test them. One might say that they are thereby asking questions to which we all know the answers. Being acquainted with the material world, we have no doubt that any two instruments of the same kind, even of the highest quality and possessing Bureau of Standards certificates, would be found to differ in their readings of the same thing if we tested them very minutely. Being acquainted with human beings, we need not stipulate very minute testing in order to reveal the differences.

What we ought to be concerned about in such studies of differences is not statistical significance but practical significance, i.e., the importance of the differences in the work that the observers and the instruments are to perform. How much variation and bias can these differences introduce, and particularly how will they compare in magnitude with other variation, e.g., between subjects? Obviously, the closer we can come to exploring these questions under actual working conditions the better. In clinical research this may be difficult; but it is very unsafe to trust the transference to a working situation of the results obtained from a study conducted under artificial conditions. Such a study may be useful as a first step, but no more.