THE USE OF THE CONCEPT OF POWER IN AGRICULTURAL EXPERIMENTATION*

By Jerzy Neyman

Research Professor, Institute for Basic Research in Science
University of California, Berkeley, California

Your Excellencies, Dr. Panse, Ladies and Gentlemen,

May I be allowed to begin by expressing my deep gratitude for the honor of being requested to address you at this opening session of your meeting. Your Society, young as it is compared to some of the sister societies in Western countries, is one of the foremost in furthering statistical research, both as an independent discipline and as an important tool in the broad field of agricultural research. Much has been already accomplished. Much more may be confidently expected in the future. Please accept my best wishes.

I have selected the power of a statistical test as the subject of to-day's talk because of the importance of this concept in the general field of experimentation and because, due to certain preoccupations with other problems, many of the contemporary experimenters, lose sight of the consideration of power and, thereby, are likely to deprive their experiments of important scientific benefits.

The power of a statistical test is one of the basic concepts of statistical theory. A test of a statistical hypothesis H consists of a rule, so-called rule of inductive behaviour, of rejecting the hypothesis H if the observations fall into a specified category, called the "critical region" w, and of abstaining from rejecting H in all other cases. (For the sake of brevity, instead of saying "abstain from rejecting H," we say "accept H".) The practical application of any such rule may result in errors of two different kinds. First, it may happen that the hypothesis under test is true and, through the unavoidable sampling variation, the sample point falls into the critical region, thus leading to the rejection of H. This particular error is conventionally called the error of the first kind. The hypothesis H is so defined that the error of the first kind is more important to avoid than the error of the second kind, to be defined presently.

The error of the second kind can be committed only if the hypothesis H is false and an alternative hypothesis h happens to be

*Delivered on 20th January 1957 on the occasion of the Tenth Annual Meeting of the Indian Society of Agricultural Statistics, at Lucknow.

true. The error of the second kind consists, then in our failure to reject H. Naturally, the importance of an error of the second kind depends on the degree of falsehood of the hypothesis tested, that is, on the difference between the hypothesis tested H and the hypothesis h that happens to be true.

Because of the postulated predominant importance of errors of the first kind, statistical tests are arranged so as to reduce the probability of a first kind error to a low level α selected by the experimenter. The established term to designate this quantity is level of significance. Any well-designed test can be applied at any chosen level of significance which, then, is totally at the disposal of the experimenter.

When a test procedure, as well as the level of significance, is fixed, the probabilities of committing errors of the second kind are perfectly determined. The term *power* is used to designate the complement to unity of the probability of an error of the second kind. In other words, the power of a given test, ordinarily designated by $\beta(h|w)$, is the probability of rejecting the hypothesis H when, in fact, this hypothesis is false and the true hypothesis is h. Thus, the power of a test, based on a specified critical region w, is a function of the hypothesis h that, in any particular case, is the true hypothesis. Another convenient way of describing the same concept is that the power $\beta(h|w)$ is the probability that the test based on w will detect the falsehood of the hypothesis tested H when the true hypothesis is h.

I wish to call your attention to the use of the concept of the power of a test in three important phases of scientific research: (i) choice of a statistical test, (ii) design of an experiment and (iii) interpretation of results.

(i) *Choice of a statistical test.*—This point may best be explained with reference to a particular study. While in Calcutta, I became acquainted with an important experiment conducted in the All-India Institute of Health under the statistical direction of Dr. C. Chandrasekharan. The experiment studies whether a particular chemical, taken by women once a month at a certain moment of the menstruation period, has a contraceptive effect. One hundred per cent. effectiveness is not expected, but it is important to achieve at least a substantial decrease in the probability of conception. At the same time, the question arises whether the effect of the chemical in question is cumulative in character. In other words, whether a decrease in the probability of pregnancy following the administration of the first pill the month of treatment, is succeeded by an additional decr same probability following the second dose of the che

In this case, the hypothesis H to be tested is that the chemical has no cumulative effect, that is, that the probability of conception is the same, no matter whether during the first period of treatment, the second, the third, etc. The alternative hypotheses are that the probabilities of conception in successive periods of treatment are different, perhaps steadily decreasing.

Superficially, one might perhaps think that the appropriate test of the hypothesis H just defined should consist in a rule of rejecting H whenever the frequencies of conception during the several consecutive periods of treatment show a substantial variation. For example, one might think of the familiar χ^2 test for homogeneity. However, upon examining the situation more closely, it is seen that a straight application of the χ^2 test is totally inadequate. The reason is that, according to the women's age, health and other circumstances, the individual probability of pregnancy during a period varies from one woman to the next. In fact, some women have difficulties in becoming pregnant even under artificial insemination and others conceive in spite of considerable care in taking precautions. Thus the whole population of women can be divided into a number of categories, each with a different probability of conception, say, $p_1 < p_2 < \cdots < p_s$. For the sake of simplicity, assume that there are only two such categories with $p_1 = \frac{1}{4}$ and $p_2 = \frac{3}{4}$. Furthermore, imagine that these two categories are equally numerous. In these conditions, the expected number of conceptions during the first monthly period of observation of some $N = 200$ randomly selected women is 25 per cent. for the first category and 75 per cent. for the second, or 50 per cent. for the whole sample. The women who become pregnant during the first month are thus withdrawn from the experiment. Now, for the second month of observation, the first category of women is expected to include 75 women and the second category only 25. The expected pregnancies in the second period will be $75/4 = 18{\cdot}75$ and $25 \times \frac{3}{4} = 18{\cdot}75$, totalling $37{\cdot}5$ for the 100 women under consideration. Thus, under the assumption of no change in the probability of pregnancy from one period to the next, the proportion of women conceiving decreases from the 50 per cent. for the first period to $37{\cdot}5$ per cent. for the second. This is the unavoidable effect of selection. We have verified this effect for the simplest case of two equally numerous categories of women with two different individual probabilities of conception, but it must be clear that similar selection effect must be present in any real study. As a result, the χ^2 test applied to frequencies of conception in successive periods of treatment tends to indicate heterogeneity which must be present even in the total absence of a cumulative effect of the treatment.

The question now arises as to what other test can be used to differentiate between the selection effect and the possible cumulative effect of the contraceptive chemical. The answer is: arrange your test so that (*a*) if there is no cumulative effect of the treatment then the probability of asserting that this effect exists is no more than the level of significance α adopted, and (*b*) among all possible tests satisfying (*a*) select the one for which the power of detecting the falsehood of the hypothesis tested is the most satisfactory.

As is well known, the mathematical side of the solution indicated frequently is not easy. However, the important point is that a rational selection of an appropriate test is impossible unless it is based on considerations of power.

(ii) *Role of power in the design of experiments.*—If we go to the trouble of setting up an experiment this is because we want to establish the presence of some possible effect of a treatment. In these circumstances it is expedient, prior to the actual performance of the experiment, to visualize the chances of detecting the contemplated effect if this effect exists and is of an important magnitude. The simplest illustration of the point in question is provided by experiments contemplated to study the frequency of success of a treatment evaluated on a "Yes" or "No" basis. Denote by n the number of independent replicates. Each replicate can indicate either a success or a failure of the treatment. The question is whether the probability p of success is no more than one-half (treatment has no effect or is harmful) or greater (treatment is beneficial). Let X denote the number of successes to be observed. The hypothesis tested H is the $p \leqq \frac{1}{2}$. The alternatives are $p > \frac{1}{2}$. The obvious criterion for testing H is X, the number of successes. If X be large, we would reject H and act on the assumption that the treatment is beneficial, but not otherwise. If the treatment is either harmful or ineffective, then the adoption of the attitude that it is beneficial is obviously undesirable. Therefore, we control this kind of error (the error of the first kind) by requiring that, for our recognition of the treatment as beneficial, the value of X be at least equal to a limit $k(n, \alpha)$ so chosen that, should the treatment be harmful or ineffective the probability of observing $X \geq k(n, \alpha)$ does not exceed a preassigned small limit α, say $\alpha = \cdot 01$. The limit $k(n, \alpha)$ may be obtained by consulting the available tables of the binomial distribution.

All this is a very usual and perfectly routine procedure. The currently non-routine question to which I wish to call your attention is concerned with the power of the test. Remember: The experiment discussed is meant to detect the beneficial effect of the treatment if

such exists. If it does exist then $p > \frac{1}{2}$. Naturally, should the value of p be greater than one-half but only slightly so, say $p = \cdot 500001$, the effect of treatment, though beneficial, is so small that it is not worth thinking about. It is a practical and subjective question as to what value of p, greater than one-half, represents an important advantage of the given treatment. Suppose an experimenter considers that the value of the probability $p = \cdot 6$ already represents an important advantage. If so, then it is important that the design of the experiment ensure a reasonable chance of detecting the beneficial effect of the treatment if its intensity is as large as $p = \cdot 6$ or greater.

A direct answer to the question of how frequently the contemplated test, applied at a level of significance α, will detect the advantage of the treatment if this advantage has a specified intensity, can be obtained from the tables of the binomial distribution. Currently there are several such tables available. Table I has been compiled using the "Tables of the Cumulative Binomial Probability Distribution" published by the Harvard University Press, 1955.

TABLE I

Critical values k *of the binomial variable, probabilities* P_1 *of errors of first kind and power* β (p) *of the test of the hypothesis that* $p = \frac{1}{2}$

n	$k(\cdot 05, n)$	P_1	$\beta(\cdot 6)$	$\beta(\cdot 7)$	$k(\cdot 01, n)$	P_1	$\beta(\cdot 6)$	$\beta(\cdot 7)$
10	8	·055	·167	·483	9	·011	·046	·149
15	11	·059	·217	·515	12	·018	·090	·297
20	14	·057	·250	·608	16	·006	·051	·238
25	17	·054	·274	·677	19	·007	·074	·341
30	20	·049	·291	·730	22	·008	·094	·432
35	23	·045	·305	·770	25	·008	·112	·510
40	26	·040	·317	·807	28	·008	·129	·577

Table I is divided into two parts of four columns each. Following the column giving the contemplated number n of replicates in the experiment, there is a column of the critical values $k(\cdot 05, n)$ such that,

if $X \geq k$ ($\cdot 05$, n) the hypothesis will be rejected at the intended (approximate) level of significance $\alpha = \cdot 05$. The next column gives the actual values of the probability P_1 of rejecting the hypothesis tested when $p = \cdot 5$. It will be seen that the actual values of P_1 differ but little from $\cdot 05$. The two following columns give the power of the test corresponding to alternative values of p, namely $p = \cdot 6$ and $\cdot 7$. The remaining four columns of Table I give similar figures corresponding to the intended (approximate) level of significance $\alpha = \cdot 01$.

In the design of the contemplated experiment there are just two elements that are left undetermined. They are the number n of replicates and the level of significance α at which to apply the test. In far too frequent cases, both relating to simplest experiments of the kind discussed and to those of much more complicated nature, questions about the number of replicates and about the level of significance are answered carelessly and offhandedly. With reference to a simple design like the one used here for illustration, one might hear the answer: The value $\cdot 05$ is too lenient to indicate real significance: let us use $\alpha = \cdot 01$. As to the number of replicates, we may be told that it should not be too small so that significance could be established without requiring that *all* the experiments result in a success. This advice is occasionally followed by the remark: Do not be afraid of small samples. With $n = 15$, the approximate level of significance $\cdot 01$ (actually $\cdot 018$) can be assured with three failures in the total experiment. This, then, may be a reasonable size of the experiment, $n = 15$.

This is typical advice that one frequently hears. Table I indicates that this advice is thoughtless. In order to see this, let us examine the power of the test corresponding to $\alpha = \cdot 01$, with $n = 15$ and to the alternative hypotheses $p = \cdot 6$ and $p = \cdot 7$. We find $\beta(\cdot 6) = \cdot 090$ and $\beta(\cdot 7) = \cdot 297$. This means that, if the experiment is performed with $n = 15$ replications and if one uses the (approximate) level of significance $\alpha = \cdot 01$, and if the hypothesis $p = \cdot 5$ is wrong, with the true value of p being either $\cdot 6$ or $\cdot 7$, then the chance that the test will discover the beneficial effect of the treatment is either nine in a hundred or 297 in a thousand, respectively. Should the experimenter consider that the probabilities of success of the treatment amounting to $\cdot 6$ or $\cdot 7$ are of importance, then the experiment ensuring the probabilities of discovering the advantage of the treatment amounting to something less than one in three, is a poor experiment.

Table I suggests two conclusions. The first is that, if $p = \cdot 6$ does represent an important advantage of the treatment, then the size of the experiment should be substantially increased beyond the range

of the table. The other conclusion is that, if one contemplates the value of $p = \cdot 7$ and if a further increase in the size of the experiment is embarrassing, then there may be an advantage in changing the level of significance, perhaps from $\alpha = \cdot 01$, to $\alpha = \cdot 05$. With this level of significance and with $n = 40$, the probability of detecting the fact that the treatment is advantageous when, in actual fact, $p = \cdot 7$, is equal to $\cdot 807$.

As already suggested, the relative importance of this or that occurrence is a highly subjective matter and it is possible that, in some cases, some experimenters will feel that a treatment represents a real advantage only if this treatment succeeds in, say, 99 per cent. of the cases. Then the relevant value of the power of the test will be that corresponding to $p = \cdot 99$, and, in such a case, the level of significance $\alpha = \cdot 01$ and the value of $n = 15$ (or smaller) may be found satisfactory. The point I wish to make is that the choice of the value of α and of the size n of the experiment should be made not blindly but after examining a table of the power of the test, similar to my Table I. In fact, it is safe to say that if experimenters realized how little is the chance of their experiments discovering what they are intended to discover, then a very substantial proportion of the experiments that are now in progress would have been abandoned in favour of an increase in size of the remaining experiments, judged more important.

The level of significance to be used and also the size of the experiment are certainly important elements of a design. However, in cases of an experiment more complicated than the one discussed, there enter many other elements. For successful experimentation it is important to evaluate all the elements of the experimental design with reference to the power of the proposed test. In other words, before settling on a particular design, it is essential to be clear about the probability that the experiment will discover what it is intended to discover, given that the effect sought exists and has an important intensity.

(iii) *The role of power in interpreting experimental results.*—In the preceding Section I discussed the use of the power of the test in designing an experiment. Now let us consider the situation when we are faced with the necessity of interpreting the results of an experiment planned and performed by someone else. Here again the numerical values of the power of the test may be very useful.

In perusing the literature concerned with agricultural experimentation, one frequently finds statements to the effect that "the interactions are not significant and that, therefore, the effects of particular

treatments may be judged by averaging simple effects of each treatment shown in the presence and absence of other treatments". It will be noticed that assertions of this kind are equivalent to adopting a sequence of rules including the following links: (*a*) test for significance of interactions. If these prove significant then (*b*) evaluate treatments in one particular way. If the interactions are not significant, then (*c*) evaluate the treatments in a different way. This change in the second step, either (*b*) or (*c*), according to the outcome of (*a*) is evidence of the recognition by the authors of a danger from an inappropriate choice between (*b*) and (*c*). A closer examination of the problem indicates that the danger is a very real one.[1] In fact, if interactions between the treatments exist then the evaluation of the treatments by their average effects may lead to most undesirable errors. Unfortunately, the recognition of the danger that the assessment of treatments by their average effects may be vitiated by the interactions is not usually followed by any protective measures and, in many cases, the rule of acting as if the interactions were non-existent whenever they are not found significant is followed automatically and blindly.

The protection against the danger of the errors indicated is in the use of the tables of the power of one analysis of variance test.[2] When the interactions are found not significant and thus step (*c*) comes under consideration, it is appropriate to do a little calculation and to attempt to answer the following questions.

(α) How large must the hypothetical interaction be in order to cause a serious error in the interpretation of the results?

(β) What was the probability (power) of detecting interactions of this particular size in the experiment performed?

Regrettably, when studying the reports on a number of agricultural trials, I frequently find unsatisfactory answers to the above question (β). The probability in question is frequently relatively low, of the order of one-half. It must be obvious that in cases of this kind the fact that the test failed to detect the existence of interactions does not mean very much. In fact, they may exist and have gone undetected.

Whatever I said about interactions applies with equal force to other cases when a test failed to detect a significant effect. In cases of this kind, to act as if the hypothesis tested has been established is obviously precipitous. A partial protection against errors is the calculation of the power of the test. Should this calculation show that the probability of detecting an appreciable error in the hypothesis tested was large, say ·95 or greater, then and only then is the decision in favour

of the hypothesis tested justifiable in the same sense as the decision against this hypothesis is justifiable when an appropriate test rejects it at a chosen level of significance.

REFERENCES

Neyman, J. .. "Contribution to the discussion of the paper by F. Yates," *Suppl. J. Roy. Stat. Soc.*, 1935, **2**, 235–41.

Pearson, E. S. and Hartley, H. O. *Biometrika Tables for Statisticians*, 1954.