The testing of statistical hypotheses in relation to probabilities a priori. By J. NEYMAN, Lecturer at the Central College of Agriculture, Warsaw, and E. S. PEARSON, Department of Applied Statistics, University College, London. (Communicated by Mr G. U. YULE.)

[*Received* 31 May, *read* 30 October 1933.]

1. *Introduction.*

In a recent paper* we have discussed certain general principles underlying the determination of the most efficient tests of statistical hypotheses, but the method of approach did not involve any detailed consideration of the question of *a priori* probability. We propose now to consider more fully the bearing of the earlier results on this question and in particular to discuss what statements of value to the statistician in reaching his final judgment can be made from an analysis of observed data, which would not be modified by any change in the probabilities *a priori*. In dealing with the problem of statistical estimation, R. A. Fisher has shown how, under certain conditions, what may be described as rules of behaviour can be employed which will lead to results independent of these probabilities; in this connection he has discussed the important conception of what he terms fiducial limits†. But the testing of statistical hypotheses cannot be treated as a problem in estimation, and it is necessary to discuss afresh in what sense tests can be employed which are independent of *a priori* probability laws.

The following brief statement summarises the position from which we start:

(1) The observed event can be represented by a point Σ, which we shall call the sample point, in a space of n dimensions, having coordinates

$$x_1, x_2, \ldots, x_n. \qquad (1)$$

The complete space within which it is permissible for Σ to lie will be termed the sample space W.

(2) H_0, the hypothesis to be tested, concerns the origin of the event, and is such as to determine the probability of occurrence

$$p_0 = p_0(x_1, x_2, \ldots, x_n) \qquad (2)$$

of every possible event. H_0 may be simple (p_0 completely specified) or composite (p_0 only partly specified). In statistical terminology Σ may represent a sample or samples, and H_0 may be a hypothesis concerning the population from which they have been drawn.

* Neyman and Pearson, *Phil. Trans. Roy. Soc.* A, 231 (1933), 289.

† Fisher, *Proc. Camb. Phil. Soc.* 26 (1930), 528; *Proc. Roy. Soc.* A, 139 (1933), 343.

(3) In any given problem it is assumed possible to define the class of admissible hypothesis $C(H)$, containing $H_1, H_2, \ldots$ as alternatives to H_0.

(4) As a result of the statistical analysis we shall decide either (a) to accept H_0, (b) to reject H_0, (c) to remain in doubt on the grounds that the evidence provided by the data is inadequate.

(5) A statistical test is therefore equivalent to a rule of the following type,

(a) Reject H_0 if Σ falls into a region w.

(b) Accept H_0 if Σ falls into another region w'.

(c) Remain in doubt if Σ falls into a third region w''.

In the following treatment we shall consider only the question of a single division into (a) the *critical region* w, (b) the *region of acceptance*, $\overline{w} = w' + w'' = W - w$. The region of doubt may be obtained by a further subdivision of the region of acceptance, but it is primarily important to discuss the consequences of a single division into two regions.

(6) In making decisions (a) or (b) we shall sometimes be in error, for problems are rarely presented in such a form that we can discriminate with certainty between the true and the false hypothesis. These errors will be of two kinds:

(I) we reject H_0 when it is true,

(II) we accept H_0 when some alternative H_i is true.

The problem before us is to consider how these errors may be best controlled. It is evident at once not only that additional information beyond that supplied by the observations is necessary in order to define $C(H)$, but also that the probabilities *a priori* ϕ_i of the admissible hypotheses $H_i (i = 0, 1, 2, \ldots)$ falling within $C(H)$ must enter into the problem. (For example, the chance of a type I error will decrease as ϕ_0 decreases.) Yet if it is important to take into account probabilities *a priori* in drawing a final inference from the observations, the practical statistician is nevertheless forced to recognize that the values of ϕ_i can only rarely be expressed in precise numerical form. It is therefore inevitable from the practical point of view that he should consider in what sense, if any, tests can be employed which are independent of probabilities *a priori*. Further, the statistical aspect of the problem will appeal to him*. If he makes repeated use of the same statistical tools when faced with a similar set of admissible hypotheses, in what

* This aspect of the error problem is very evident in a number of fields where tests must be used in a routine manner, and errors of judgment lead to waste of energy or financial loss. Such is the case in sampling inspection problems in mass-production industry.

sense can he be sure of certain long run results? A certain proportion of correct decisions, a certain proportion of errors, and if he so formulates the tests a certain proportion of cases left in doubt?

It will be noticed that we have assumed it possible to define the class $C(H)$ of admissible hypotheses H_i. In the formal treatment of the subject which follows this is no limitation, for we may set the limits of $C(H)$ as widely as we please; and if we do not know whether the probability *a priori* of meeting certain of the alternatives is zero, this will not matter since we are searching for tests which would be valid independently of this knowledge. But as soon as we attempt to make use of the formal results in practical application, it is found that some assumptions regarding $C(H)$, of a limiting nature, must generally be made; we must assume, for example, that the admissible hypotheses concern normally distributed populations*. It is not easy to see how this difficulty can be avoided, but perhaps the most hopeful method of attack is first to construct a consistent picture on simplified assumptions, and afterwards to examine the nature and extent of errors in judgment that may result if $C(H)$ is really wider than we have had to assume in order to make mathematical treatment possible. Certain investigations on these lines have been carried out from time to time.

2. *The case of a simple hypothesis.*

Suppose that the admissible hypotheses concerning the origin of the observed event are $H_0, H_1, \ldots, H_m$, and that one or other of these must be true. For simplicity we shall take the number $m+1$ of these alternatives as finite. Denote by ϕ_i the probability *a priori* that the true hypothesis is H_i $(i=0, 1, 2, \ldots, m)$. Since we have assumed that there are only $m+1$ possible alternatives, the probabilities ϕ_i must satisfy the condition

$$\sum_{i=0}^{m} (\phi_i) = 1. \qquad (3)$$

The following notation will be convenient. If A indicates an observed event and B is a hypothesis concerning its origin, then $P(A \mid B)$ will denote the probability of the event A, calculated on the assumption that B is true. For example, $P(\Sigma \mid H_0)$ will denote

* In problems dealing with grouped frequencies this may not be so, for if the chance of an individual unit falling into one of k alternative categories is

$$p_t \quad (t=1, 2, \ldots, k),$$

then $C(H)$ may be allowed to include all possible sets of non-negative p's, subject to the sole condition $\Sigma(p_t)=1$. In this way the χ^2 tests, although based on certain mathematical approximations, are of wider application than those dealing with criteria based on the symmetric functions of continuous variables.

the probability of a sample Σ, calculated on the assumption that the sampled population is accurately described by the hypothesis H_0—or in other words that H_0 is true.

If now we have chosen a region w, in the sample space W, as critical region to test H_0, then the probability that the sample point Σ defined by the set of variates (1) falls into w, if H_0 is true, may be written as

$$P(w \mid H_0) = \epsilon. \tag{4}$$

The chance of rejecting H_0 if it is true is therefore equal to ϵ, and w may be termed of size ϵ for H_0. The second type of error will be made when some alternative H_i is true, and Σ falls in $\overline{w} = W - w$. If we denote by $P_{\mathrm{I}}(w)$, $P_{\mathrm{II}}(w)$ and $P(w)$ the chance of an error of the first kind, the chance of an error of the second kind and the total chance of error in using w as critical region, then it follows that

$$P_{\mathrm{I}}(w) = \phi_0 P(w \mid H_0), \tag{5}$$

$$P_{\mathrm{II}}(w) = \sum_{i=1}^{m} \phi_i P(\overline{w} \mid H_i), \tag{6}$$

$$P(w) = P_{\mathrm{I}}(w) + P_{\mathrm{II}}(w). \tag{7}$$

We may now discuss what meaning could be given to the words: "a test independent of the probability law *a priori*."

Definition A. The phrase might be defined as implying a choice of critical region w in such a way that the probability $P(w)$ of making an error in testing H_0 had a value independent of the probabilities ϕ_i.

To ascertain whether such a region can be found we must consider the equations (5), (6) and (7). It is easy to see that a necessary and sufficient condition for $P(w)$ to be independent of ϕ_i is that

$$P(w \mid H_0) = P(\overline{w} \mid H_i) = \epsilon \qquad (i = 1, 2, \ldots, m). \tag{8}$$

When this is so, in view of (3) it will follow that

$$P(w) = \epsilon. \tag{9}$$

Before considering how far critical regions satisfying the relation (8) can in general be found, we may illustrate the result by a single example. Suppose that H_0, the hypothesis to be tested, consists of the assumption that the proportion of black balls in a bag is $p_0 = \frac{1}{4}$, and that there is only one alternative, namely that this proportion is $p_1 = \frac{3}{4}$. A sample of an odd number of balls n is randomly drawn, each ball being replaced in the bag before the next is drawn. The proportion of black balls in this sample is $x = k/n$. Clearly either

https://doi.org/10.1017/S030500410001152X Published online by Cambridge University Press

$x > \frac{1}{2}$ or $x < \frac{1}{2}$, and we may take, as the critical region for testing H_0, that defined by the inequality $x > \frac{1}{2}$. Then

$$P(w \mid H_0) = \sum_{x>\frac{1}{2}} \left\{ \frac{n!}{(nx)!\,(n-nx)!}\, p_0^{nx} p_1^{n-nx} \right\} = P(\overline{w} \mid H_1) = \epsilon, \quad (10)$$

and the chance of making an error in following this rule will be ϵ, whatever be the probabilities *a priori* of the two alternatives. As the size of sample is increased (n remaining odd), $P(w) - \epsilon$ will tend to zero, and probably no serious objection could be raised against the test.

It is possible that there may be many analogous cases in which a test independent of the probabilities *a priori* (in the sense of Definition A) could be found, but it is necessary to call attention to a very wide class of problems in which even if a region satisfying (8) could be determined, it would provide a test of no practical value. Such problems are those in which the set of alternative hypotheses are very numerous (if not infinite), and some of these differ only insignificantly from the hypothesis tested. When, for example, we test the hypothesis that a sample has been drawn from a normal population with mean a_0 and standard deviation σ_0, the hypothesis that $a = a_0 + \delta a$, $\sigma = \sigma_0 + \delta\sigma$ will generally be a permissible alternative*. In general we shall find among the alternatives one, say H_1, such that whatever be the region w, $P(w \mid H_1)$ differs but little from $P(w \mid H_0)$. Then the approximate equality

$$P(w \mid H_1) = P(w \mid H_0), \quad (11)$$

together with the exact one

$$P(\overline{w} \mid H_1) = P(w \mid H_0), \quad (12)$$

would lead to

$$P(w \mid H_0) = \tfrac{1}{2} \text{ (approximately).} \quad (13)$$

In cases where the set of alternatives is infinite and such that by a proper choice we may find a series $H^{(1)}, H^{(2)}, \ldots, H^{(k)}, \ldots$ for which

$$\lim P(w \mid H^{(k)}) = P(w \mid H_0), \quad (14)$$

whatever be the region satisfying (8), the equality (13) will be an exact one. If then we were to use such a region in the test, we should know that whatever be the probabilities *a priori* of the alternatives, the chance of an error in judgment would be equal to $\frac{1}{2}$. The test would therefore be of no value at all†.

* The question, which is clearly important, of how far it matters accepting H_0 falsely when the true hypothesis differs only slightly from it, is referred to again below.

† It is true that we might remove from the summation in (6) certain of the H_i differing only slightly from H_0, on the grounds that the consequence of accepting H_0 when they were true was not serious enough to be termed an "error"; and in this way we might find a region, w, for which ϵ was much less than $\frac{1}{2}$. But even then it is not clear that a region satisfying (8) would exist.

We think indeed that it is safe to conclude that while special cases can be found in which the Definition A would lead to a practical test, in general this interpretation of the phrase "a test independent of the probability law *a priori*" is unworkable. We must accept the fact that without a knowledge of the probabilities ϕ_i we cannot in general calculate $P(w)$, and further that without knowing ϕ_0 we cannot find the relative magnitude of the chances of the two sources of error. Certain further conceptions must however be examined before suggesting an alternative definition.

The first of these is that there is an essential difference in character between errors of type I and type II. If a true hypothesis is rejected, a mistake has been made whose consequence will generally be the same whatever be the sample to which the test was applied. That is to say all errors of type I may be described as equivalent. On the other hand if H_0 is accepted when some alternative H_i is true, the consequences that follow will depend upon the nature of H_i and its difference from H_0. In rejecting a true hypothesis we have closed for the time being the route to our goal, but have not necessarily determined what other route to take; in accepting a hypothesis which is not true we have taken a route which may differ only very slightly but may also differ very much from the true one for which we are searching. Generally it will not be of serious consequence if we accept H_0 falsely when the true hypothesis H_i only differs from it slightly, but the danger will steadily increase as this difference increases. It follows that in discriminating between alternative critical regions w, while it is the "size" of $P_{\mathrm{I}}(w)$ that matters, it is what may be termed the quality of the errors contributing to the risk, $P_{\mathrm{II}}(w)$, that must be taken into account.

It is also evident that in certain cases we attempt to adjust the balance between the risks P_{I} and P_{II} to meet the type of problem before us. Some examples will illustrate this point.

In a scientific investigation we may be testing some new hypothesis H_0, and for this purpose comparing the number of observations, $n_1, n_2, \ldots$, falling into certain groups with theoretical values $m_1, m_2, \ldots$ derived from H_0. The hypothesis is perhaps novel and important, and we do not wish to throw it aside lightly. In applying the χ^2 test to the frequencies n_s and m_s we shall therefore be inclined to give H_0 the benefit of the doubt, and fix the level of rejection low so that

$$P(\chi^2 \geqslant \chi_0^2) = P(w \mid H_0) = \epsilon \qquad (15)$$

is perhaps ·01 or less. (χ_0^2 is here supposed to be the observed value.)

On the other hand we may be analysing the results of a series of experiments designed to detect possible factors which may modify

https://doi.org/10.1017/S030500410001152X Published online by Cambridge University Press

the working of a standard law. In this case we shall be watching carefully for any signs of divergence from the standard hypothesis H_0, and shall allow ϵ to be large—perhaps ·10—in order that the risk of error II may be reduced. The importance of finding some new line of development here outweighs any loss due to a certain waste of effort in starting on a false trail.

An example of a different kind arises in connection with the sampling tests laid down in commercial specifications. H_0 is the hypothesis that the consignment which is sampled is of quality above a certain standard*. From the *producer's* point of view it is important that the sample should not be rejected when H_0 is true; he wishes P_{I} to be reduced to a minimum. To the *consumer* on the other hand it is important that the sample should not pass the test when H_0 is false, the quality of the consignment being below standard; his object will be to reduce P_{II}. At the same time the stringency of the testing rule which the consumer employs would rightly depend upon the confidence that he places in the producers with whom he deals—that is to say on the values of ϕ_i.

Bearing these points in mind, let us again consider what statements, likely to be of practical value, can be made that are independent of probabilities *a priori*. Certain further definitions will be necessary.

(*a*) We shall describe two different tests T_1 and T_2, associated with critical regions w_1 and w_2, as *equivalent* when the probabilities $P_{\text{I}}(w_1)$ and $P_{\text{I}}(w_2)$ of making an error of type I are equal. Since we are dealing with a simple hypothesis H_0, this implies the relation

$$P(w_1 \mid H_0) = P(w_2 \mid H_0) = \epsilon. \tag{16}$$

(*b*) Consider a critical region w of size ϵ for H_0, and let H_i be an alternative simple hypothesis. The probability of rejecting the hypothesis tested, H_0, when the true hypothesis is H_i, or $P(w \mid H_i)$, may be termed the *power* of the critical region w with regard to H_i. Thus, if for any other region w_1

$$P(w_1 \mid H_i) < P(w \mid H_i), \tag{17}$$

we shall say that the test T based on w is more powerful with regard to H_i than the test T_1 based on w_1. Clearly a test which is more powerful with regard to one alternative H_i than is a second test T_1 may be less powerful with regard to some other alternative H_j, so that both (17) and

$$P(w_1 \mid H_j) > P(w \mid H_j) \tag{18}$$

may be true. It may happen, however, that whatever be the alternative H_i belonging to the class $C(H)$, the test T is more

* It would be a composite and not a simple hypothesis.

https://doi.org/10.1017/S030500410001152X Published online by Cambridge University Press

powerful than T_1 with regard to H_i. If this be so, we shall describe T as *uniformly more powerful* than T_1 with regard to the class of alternatives $C(H)$.

(*c*) If now we consider the probability $P_{II}(w)$ of type II errors when using a test T based on the critical region w, we may describe

$$1 - \phi_0 - P_{II}(w) = \sum_{i=1}^{m} \{\phi_i P(w|H_i)\} \qquad (19)$$

as the *resultant power* of the test T. Thus for a given ϕ_0, the greater the resultant power, the smaller will be $P_{II}(w)$. It is seen that while the power of a test with regard to a given alternative H_i is independent of the probabilities *a priori*, and is therefore known precisely as soon as H_i and w are specified, this is not the case with the resultant power, which is a function of the ϕ_i's.

It is difficult to compare tests which are neither equivalent nor of equal resultant power. Generally, however, a "test" does not consist of a single critical region, but of a family of regions each corresponding to different values of $P(w|H_0) = \epsilon$. A comparison between T_1 and T_2 may then be made by choosing a region from each family such that the resulting tests are equivalent. From this would follow:

Definition B. A test T will be termed independent of the probabilities a priori ϕ_i *if it is of greater resultant power than any other equivalent test, whatever be these probabilities.*

It will be noticed that the words "independent of the probabilities *a priori*" are not here used in the sense that changes in the probabilities ϕ_i would not modify the chance of error in testing the hypothesis; they would do so. What the definition implies is that even with a knowledge of the ϕ_i's we could not better our choice of a critical region as far as equivalent tests are concerned.

We shall now show:

(1) that if a test T satisfies the Definition B, then it is uniformly more powerful with regard to $C(H)$ than any other equivalent test;

(2) and, vice versa, if a test T is uniformly more powerful with regard to $C(H)$ than any other equivalent test, then it will satisfy Definition B.

Assume that the test T_1 associated with the critical region w_1 satisfies Definition B, and consider some other equivalent test T_2 based on the critical region w_2. Equation (16) will hold, and further the difference

$$P_{II}(w_2) - P_{II}(w_1) = \sum_{i=1}^{m} \phi_i \{P(w_1|H_i) - P(w_2|H_i)\} \qquad (20)$$

will be positive (or zero) whatever the probabilities ϕ_i. This is possible only if

$$P(w_1|H_i) - P(w_2|H_i) \geqslant 0 \qquad (i = 1, 2, \ldots, m). \qquad (21)$$

Otherwise, if (21) was not satisfied, let us say, for $i = 1$, then, if

$$\phi_1 = 1, \quad \phi_i = 0 \qquad (i \neq 1), \qquad (22)$$

(20) would be negative and T_1 would not satisfy Definition B, since for the system of ϕ_i's defined by (22) it would be of smaller resultant power than T_2. Thus, if a test T_1 satisfies Definition B, (21) must hold whatever the alternative H_i and whatever the region w_2 equivalent to w_1.

On the other hand if (21) holds, whatever the alternative H_i and whatever the region w_2 equivalent to w_1, in other words if T_1 is uniformly more powerful with regard to the class $C(H)$ than any other equivalent test, then the difference (20) cannot be negative, and the test T_1 satisfies Definition B.

We see then that the problem of finding tests satisfying Definition B is that of finding tests which are uniformly more powerful with regard to the class $C(H)$ than any other equivalent test, and this problem may be restated as follows:

To find a region w_0 in the sample space such that

(a) $$P(w_0|H_0) = \epsilon, \qquad (23)$$

(b) if w_1 is any other region whatsoever for which

$$P(w_1|H_0) = \epsilon, \qquad (24)$$

then $$P(w_0|H_i) \geqslant P(w_1|H_i) \qquad (25)$$

for every admissible alternative hypothesis H_i.

This problem we have studied elsewhere*, and shown that it may be solved in many cases met with in practice. There are, however, cases where no solution exists, namely where the region w_0 which maximises $P(w|H_i)$ for a certain alternative H_i will not do so for another alternative. The regions maximising $P(w|H_i)$ under the condition (23) were described in the paper referred to as *best critical regions* for the hypothesis H_0 with regard to the hypothesis H_i. If a region possesses this property for every hypothesis H_i belonging to a certain class $C(H)$, then it is said to be a best critical region with regard to the whole class. We may say now that if tests exist at all, independent of the probabilities *a priori*, in the sense of Definition B, then they are associated with best critical regions.

Since such tests are independent of the probabilities *a priori*, in the sense of Definition B, they minimise the risk of type II errors

* Neyman and Pearson, *Phil. Trans. Roy. Soc.* Series A, 231 (1933), 289–337.

for every alternative hypothesis, and consequently no considerations of the different qualities of these errors need trouble us. At the same time, although we are unable to determine the proportion of type I errors that we shall make in our statistical experience as long as ϕ_0 is unknown (since $P_I(w_0) = \phi_0 \times \epsilon$), we can control the chance of rejecting a true hypothesis when we meet it (since $P(w_0|H_0) = \epsilon$), and we know that $P_I(w_0)$ has ϵ for its upper bound.

If there is no common best critical region with regard to the class $C(H)$, we cannot find a test satisfying Definition B. It would seem that some method of scaling or classification of the different type II errors is now required. In cases where Definition B is applicable, this has been unnecessary owing to the property of uniform power associated with tests based on best critical regions. Considerations which must be taken into account when this property is absent may be illustrated by an example.

Let H_i be the hypothesis that a sample of n values of a single variable has been drawn from a normally distributed population with mean a_i, and standard deviation σ_i. Suppose that H_0, the hypothesis to be tested, is that

$$a_0 = 400, \qquad \sigma_0 = 50$$

and that the admissible alternatives can be classed into four groups of normal populations, the means and standard deviations of the populations within a group differing by not more than a unit or two*. Let these groups be typified by central values as follows:

Group (1)	mean	397,	standard deviation	48.
Group (2)	„	380,	„	„ 42.
Group (3)	„	420,	„	„ 58.
Group (4)	„	500,	„	„ 70.

The occurrence of type II errors associated with the different groups (i.e. of accepting H_0 when an alternative belonging to the group is really true) will have different consequences which we will suppose summarised as follows:

Group (1), not at all serious; groups (2) and (3), moderate and of about equal importance; group (4), serious.

Let ϕ_1, ϕ_2, ϕ_3, and ϕ_4 be the probabilities *a priori* of meeting in our experience with a sample from a population belonging to group (1), (2), (3), and (4) respectively. Then *if the ϕ_i's were known* the following considerations could hardly be left out of account in determining a critical region which satisfies our intuitional requirements:

(1) An infinite variety of equivalent critical regions can be determined, for which $P(w|H_0) = \epsilon$.

* We have shown (*loc. cit.*) that in this case there is no common best critical region, i.e. no test satisfying Definition B.

https://doi.org/10.1017/S030500410001152X Published online by Cambridge University Press

(2) In our choice from among these we shall not be much influenced by the risk of type II errors associated with group (1), unless ϕ_2, ϕ_3 and ϕ_4 are all negligibly small.

(3) If ϕ_3 and ϕ_4 were both exceedingly small, say less than ·0001, but ϕ_2 was not small, we should determine the critical region so as to reduce to a minimum the risk of accepting H_0 when an alternative from group (2) was really true.

(4) On the other hand if ϕ_4 was still less than ·0001, but both ϕ_2 and ϕ_3 were large, we should take into account both these groups.

(5) If ϕ_4 was not small, we should be primarily concerned in fixing the critical region so as to prevent type II errors associated with group (4).

If it is agreed that considerations such as these are relevant when the ϕ_i's are known, then it is clear that, when they are unknown, any test that we may adopt will not be independent of the probabilities *a priori* in the sense defined above.

The problem is not easy and it is likely that no solution can be found having the unique status which seems to belong to tests associated with best critical regions. The following methods of treatment are suggested:

(*a*) We have discussed elsewhere* the use of a criterion based on a maximum likelihood ratio to obtain a critical region which eliminates type II errors with increasing stringency as H_i differs more and more from H_0. Such a region is fixed without any reference to probabilities *a priori*, and the resulting test cannot be described as independent of these probabilities, since, if the ϕ_i's were known, a test of greater resultant power could almost certainly be found. The method seems, however, to lead to useful compromise regions, and in the cases which we have examined the tests appear to satisfy our intuitional requirements.

(*b*) Problems may occur in which all type II errors can be regarded as of the same consequence, or perhaps, where these errors can be divided into two classes—those which do not matter and those which do—all the latter being treated as of equal consequence.

(*c*) Sometimes it may happen that numerical measures can be assigned to the consequences of both types of error. Such might be the case in certain industrial problems in which an error in testing a hypothesis can be associated with a financial loss, for example $£a_0$ resulting from the rejection of H_0 when it is true, and $£a_i$ from the acceptance of H_0 when H_i is true. If this was

* Neyman and Pearson, *Biometrika*, 20 A (1928), 175 and 263; *Bull. Acad. Polonaise Sci. Lettres*, Série A (1930), p. 73.

so, by introducing factors a_i we could substitute for $P_{\mathrm{I}}(w)$ and $P_{\mathrm{II}}(w)$ the mathematical expectations of losses associated with the two types of error and with the use of a particular critical region w, namely

$$\left.\begin{aligned} E_{\mathrm{I}}(w) &= a_0 P_{\mathrm{I}}(w) = a_0 \phi_0 P(w \mid H_0), \\ E_{\mathrm{II}}(w) &= \sum_{i=1}^{m} \{a_i \phi_i P(\overline{w} \mid H_i)\}. \end{aligned}\right\} \tag{26}$$

By considering these expressions, we could construct tests "independent of the probabilities *a priori*" in the sense defined above (or as defined below), with the only difference, that instead of the probabilities $P(w \mid H_0)$ and $P(\overline{w} \mid H_i)$ we must substitute the products $a_0 P(w \mid H_0)$ and $a_i P(\overline{w} \mid H_i)$. It follows that this situation (*c*) would not introduce any new principle essentially modifying the problem, although, of course, there might be considerable difference in the details of solution and its difficulty.

If the situation can be represented as in (*b*), the following method of choice of a critical region might be adopted, which would lead to a further definition of independence.

The probability of type II errors associated with the use of a critical region w has been defined by

$$P_{\mathrm{II}}(w) = \sum_{i=1}^{m} \{\phi_i P(\overline{w} \mid H_i)\}. \tag{27}$$

Denote by $\Pi(w)$ the upper bound of the probabilities $P(\overline{w} \mid H_i)$.

Clearly
$$P_{\mathrm{II}}(w) \leqslant \Pi(w), \tag{28}$$

whatever be the probabilities ϕ_i. Choose now from all equivalent regions of size ϵ for H_0, that region w_0 for which $\Pi(w_0)$ is a minimum †. If we accept the test based on w_0, it has the following properties:

(*a*) It is of size ϵ, i.e. $P(w_0 \mid H_0) = \epsilon$.

(*b*) For every other critical region w_1, equivalent to w_0, the upper bound of the probability of type II errors, namely $\Pi(w_1)$, is larger than $\Pi(w_0)$. That is to say, while w_0 and w_1 would provide equal control of type I errors, the control of type II errors is in general not the same, and may be described as follows:

(i) There is no guarantee that the test based on w_0 will be always of greater resultant power than that based on w_1; for there may be systems of the values of the ϕ_i's for which the use of w_1 will provide the better control of type II errors.

(ii) But it may be asserted that whatever the probabilities *a priori*, the probability of type II errors, $P_{\mathrm{II}}(w_0)$, when using w_0

† It is assumed that a region w_0, minimising $\Pi(w_0)$, exists.

https://doi.org/10.1017/S030500410001152X Published online by Cambridge University Press

never surpasses $\Pi(w_0)$; while there will be systems of ϕ_i's such that this probability of error, $P_{II}(w_1)$, corresponding to the region w_1 is as close to $\Pi(w_1)$ as desired, and thus is larger than

$$\Pi(w_0) \geqslant P_{II}(w_0).$$

This property of the region w_0 described in (ii), providing a known upper limit to $P_{II}(w_0)$ whatever the ϕ_i's may be, could be taken as the basis of a further definition of a test independent of the *a priori* probability law.

Definition C. A test T based on a critical region w_0 will be termed independent of the probabilities a priori, ϕ_i, *if*

(1) *whatever the ϕ_i's may be, the probability of type II errors $P_{II}(w_0)$ never exceeds a number $\Pi(w_0)$; and*

(2) *whatever other equivalent test T_1 is taken, based on a critical region w_1, there are systems of ϕ_i's such that*

$$P_{II}(w_1) \geqslant \Pi(w_0) \geqslant P_{II}(w_0),$$

and for which the test T has thus greater resultant power than T_1.

3. *The case of composite hypotheses.*

As defined above, a composite hypothesis H_i may be regarded as consisting of a set of simple hypotheses $h_{i1}, h_{i2}, \ldots, h_{ik}, \ldots$, or, if we use the terminology of mathematical logic, as the logical sum of these simple hypotheses. ϕ_{ik} will denote the probability *a priori* of h_{ik} ($i = 0, 1, 2, \ldots$; $k = 1, 2, \ldots$). Further we shall write

$$\phi_i = \sum_k (\phi_{ik}) = \phi_i \sum_k (\psi_{ik}), \tag{29}$$

where $\sum_k (\psi_{ik}) = 1$. We may now express the chance of occurrence of the first and second types of error, in testing a composite hypothesis H_0 by means of a critical region w, namely

$$P_I(w) = \sum_k \{\phi_{0k} P(w|h_{0k})\} = \phi_0 \sum_k \{\psi_{0k} P(w|h_{0k})\}, \tag{30}$$

$$P_{II}(w) = \sum_{i \geqslant 1} \sum_k \{\phi_{ik} P(\overline{w}|h_{ik})\}, \tag{31}$$

where as before $\overline{w} = W - w$ is the region for acceptance of H_0. As before we shall treat all errors of type I as equivalent, and we are therefore concerned with the value of the probability of their logical sum. From our point of approach this is an essential feature of the problem. For example, if we test the composite hypothesis that a sample has been drawn from some unspecified normally distributed population, it is only the question of normality to which we intend the test to be sensitive. To suppose that it would matter less if we rejected H_0 when the standard deviation of the normal population had some value which *a priori* we consider most

https://doi.org/10.1017/S030500410001152X Published online by Cambridge University Press

improbable, would be to misinterpret the purpose of the test as we regard it. To examine the value of the standard deviation, a separate test would need to be applied after decision on the question of normality has been taken.

If now we write

$$P(w|h_{0k}) = \epsilon_k, \tag{32}$$

we have

$$P_I(w) = \phi_0 \{\sum_k (\psi_{0k}\, \epsilon_k)\}. \tag{33}$$

As in the case of testing a simple hypothesis, we cannot determine $P_I(w)$ without knowing ϕ_0. Moreover in the present case we require also a knowledge of the ψ_{0k}'s. There are however certain statements with regard to the first source of error which can be made that are independent of the probabilities *a priori*.

(a) We may sometimes find regions w, such that $P_I(w) = \phi_0 \epsilon$, where ϵ is independent of the ψ_{0k}'s; that is to say, using such a region we shall be able to determine exactly the chance ϵ of rejecting a true hypothesis when we meet it, although ϕ_0 remains unknown. For this condition to be satisfied it follows from (33) that the region w must be such that

$$P(w|h_{0k}) = \text{constant} = \epsilon, \tag{34}$$

for all the simple hypotheses h_{0k} of which H_0 is the logical sum. Such regions we have termed *similar regions* with regard to the composite hypothesis H_0, and we have discussed the procedure for determining them if they exist*.

(b) If similar regions do not exist, we cannot find the chance of rejecting a true hypothesis when we meet it, independently of the ψ_{0k}'s, but we may be able to determine the upper bound of this chance and consequently the upper bound of $P_I(w)$. In a number of problems, one of which is suggested below, this will be adequate for our purpose. This upper bound will be the upper bound of the values of ϵ_k defined by (32).

In dealing with simple hypotheses the conception of equivalent regions was introduced. In the present case it is only possible that

$$P_I(w_1) - P_I(w_2) = \phi_0 \sum_k \psi_{0k} \{P(w_1|h_{0k}) - P(w_2|h_{0k})\} = 0 \tag{35}$$

for all values of ψ_{0k}, if

$$P(w_1|h_{0k}) = P(w_2|h_{0k}) = \epsilon_k \tag{36}$$

for all values of k. Tests for which the critical regions are such as to satisfy (36) we shall term *absolutely equivalent* with regard to the composite hypothesis H_0. Clearly tests based on similar regions

* *Phil. Trans. Roy. Soc.* A, 231 (1933), 289.

are absolutely equivalent, but they satisfy the additional condition that ϵ_k is the same for all values of k.

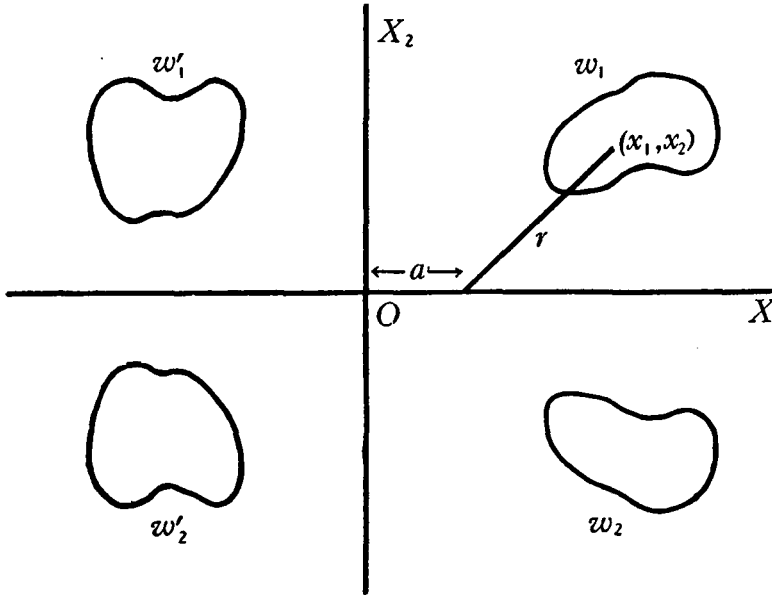
A number of sets of such absolutely equivalent regions may exist. As a simple illustration we may take a case where $n = 2$, so that every sample may be represented by a point (x_1, x_2) in a plane. Denote by $p(x_1, x_2; a)$ the elementary probability determined by a simple hypothesis h_{0a}, belonging to the set of which H_0, the hypothesis tested, is the logical sum. Suppose that the value of $p(x_1, x_2; a)$ depends only upon the distance

$$r = \{(x_1 - a)^2 + x_2^2\}^{\frac{1}{2}}$$

of the sample point (x_1, x_2) from the point $(a, 0)$ on the axis OX_1. If then we take any region w_1 and its reflection w_2 in the axis OX_1, we have

$$\iint_{w_1} p(x_1, x_2; a)\, dx_1\, dx_2 = \iint_{w_2} p(x_1, x_2; a)\, dx_1\, dx_2 = \epsilon_a, \quad (37)$$

whatever a may be. That is to say, if used for testing the hypothesis H_0, these two regions are absolutely equivalent. So also are another pair of regions w_1' and w_2' with the same property, but clearly w_1' will not be absolutely equivalent to w_1. The position is represented in the Figure. It does not, however, follow



that w_1 or w_2 are regions similar with regard to the hypothesis H_0; this would only be true if the value ϵ_a of (37) was independent of a.

https://doi.org/10.1017/S030500410001152X Published online by Cambridge University Press

With regard to the second type of error, since the alternative hypotheses consist of the set of simple hypotheses

$$h_{ik}, \qquad (i = 1, 2, \dots;\ k = 1, 2, \dots),$$

the previous definitions regarding the power of tests will be applicable without modification. We can therefore give the following definition of independence:

Definition D. A test of a composite hypothesis may be termed independent of the probabilities a priori, *if it is of greater resultant power than any other absolutely equivalent test, whatever be these probabilities.*

The following points must be noted:

(1) The region w_0 associated with the test must be such as to satisfy the conditions

$$P(w_0|h_{ik}) \geqslant P(w_1|h_{ik}) \tag{38}$$

for every other admissible simple hypothesis h_{ik}, where w_1 is any other region whatsoever, absolutely equivalent to w_0*.

(2) The test so determined is not necessarily unique, since there may be more than one set of absolutely equivalent regions. On the other hand such regions may not exist at all.

(3) So long as the values of $\phi_{0k} = \phi_0 \psi_{0k}$ are unknown, we shall only be able to obtain the upper bound of $P_I(w_0)$. In the special case when one of the sets of absolutely equivalent regions is a set of similar regions, we can however find and control the chance ϵ of rejecting a true hypothesis when we meet one.

The conditions to be satisfied are undoubtedly complicated. In the first place it may be questioned: is it ever possible that relations (34) and (38) should both be satisfied? We have dealt with this problem at length elsewhere, and shown that in a number of important statistical tests connected with samples from normal populations these conditions do hold good.† The regions associated with such tests we have termed best critical regions, as in the case of testing simple hypotheses. If it is considered essential to chose a critical region such that

$$\sum_k \{\psi_{0k} P(w_0|h_{0k})\} = \epsilon \tag{39}$$

can be given any desired value, then it appears that no better tests independent of the probabilities *a priori* than those satisfying (34) and (38) can be devised. On the other hand, similar regions for H_0 may not exist, or if they exist as well as another set of

* In other words, if a test T satisfies Definition D, it will be uniformly more powerful with regard to the class of alternatives $C(h)$ than any other equivalent test; and vice versa.

† *loc. cit.*

Such is "Student's" test, also R. A. Fisher's tests for comparing the means and variances in two samples from normal populations.

absolutely equivalent but dissimilar regions, it is still possible that we might prefer to choose a critical region from the latter. In doing this we should know only the upper bound of ϵ instead of its actual value, but we might obtain a better control of type II errors.

With regard to absolutely equivalent regions in general, it will be seen that a unique solution satisfying Definition D is possible, when

(*a*) For each simple hypothesis h_{0k} a common best critical region w_k exists with regard to each alternative hypothesis h_{ij} $(i \neq 0)$.

(*b*) When w_k is the same for all values of k, or $w_k = w_0$ (say), although $P(w_0|h_{0k}) = \epsilon_k$ is not necessarily the same.

A simple illustration of this case occurs when the hypothesis H_0 is that the mean a of a normal population with known standard deviation σ is not negative, i.e. $a \geqslant 0$, the alternative hypothesis being that $a < 0$. We have previously considered a problem closely connected with this*; namely that of determining the best critical region for testing the simple hypothesis, say h_{a_0}, that the mean of a sampled normal population of known standard deviation σ has some definite value a_0, the alternative being that $a < a_0$. It was shown that whatever the alternative, say h_a, the best critical region w is the same, namely that defined by the inequality

$$n\bar{x} = x_1 + x_2 + \ldots + x_n < k(a_0, \sigma, \epsilon), \tag{40}$$

where $k(a_0, \sigma, \epsilon)$ is a number depending only on a, σ and the value chosen for

$$P(w \mid h_{a_0}) = \epsilon. \tag{41}$$

If in dealing with the present problem we fix any value for ϵ and consider the region w_0 defined by

$$n\bar{x} = x_1 + x_2 + \ldots + x_n < k(0, \sigma, \epsilon), \tag{42}$$

we shall find that it has the following properties:

(*a*) For every simple hypothesis h_a (contained in H_0), specifying that the population mean has some definite, non-negative value a, we have

$$P(w_0 \mid h_a) = \epsilon_a \leqslant \epsilon. \tag{43}$$

(*b*) For every region w_1, which is absolutely equivalent to w_0, that is to say such that for any $a \geqslant 0$

$$P(w_0 \mid h) = P(w_1 \mid h_a) = \epsilon_a, \tag{44}$$

we have, for every $a_1 < 0$,

$$P(w_0 \mid h_{a_1}) \geqslant P(w_1 \mid h_{a_1}). \tag{45}$$

* *loc. cit.* pp. 302–304.

https://doi.org/10.1017/S030500410001152X Published online by Cambridge University Press

It follows from the preceding work that this region w_0, defined by (42), will provide a test T independent of the probabilities *a priori* in the sense of Definition D; that is to say, whatever be the values of these probabilities, T has a greater resultant power (and is also uniformly more powerful) than any other test T_1 absolutely equivalent to T. Further ϵ, which is at our choice, will be the upper bound of the chance of rejecting a true hypothesis when we meet it.

This problem is of course a simple one, and so far it has not in all cases been possible to establish that tests in common use satisfy Definition D. In cases where there is no test of a composite hypothesis satisfying this Definition, it would be possible to set down a further definition of independence analogous to that discussed in the case of testing simple hypotheses (Definition C).

4. *Conclusion.*

We began with the question: to what extent is it possible to employ tests of statistical hypotheses which are independent of probabilities *a priori*? To answer this it has been necessary to discuss what is meant by "independence". We have suggested that a statistical test may be regarded as a rule of behaviour to be applied repeatedly in our experience when faced with the same set of alternative hypotheses. From this point of view it becomes natural to analyse the errors that we shall make, which are of two types:

I. We reject the hypothesis H_0 when it is true.

II. We accept H_0 when some alternative H_i is true.

In making a decision upon which subsequent action will be based we are influenced by the consequences which follow from a wrong decision; some errors will matter more than others, and certain tests might be described as safer than others. We have therefore considered the conditions under which the choice of a test could not be improved upon—though the test might give us more information—even if the probabilities *a priori* were known.

In the first place we have suggested that all errors of type I may be regarded as equivalent, while those of type II will be of differing quality according to the extent to which the alternative hypothesis which is true varies from that which is tested. Two tests which are equivalent (H_0 simple), or absolutely equivalent (H_0 composite), assure an equal control of type I errors. A test with greater resultant power than a second ensures a smaller risk of type II errors. The following definition, which applies, with slight modification, to composite as well as to simple hypotheses, has then been discussed:

A test of a simple (composite) hypothesis may be termed independent of the probabilities *a priori*, if it is of greater resultant power than any other (absolutely) equivalent test, whatever be these probabilities.

We have shown the relation of tests satisfying these conditions to those based on what have been defined elsewhere as best critical regions. Further we have suggested other lines of attack when such regions do not exist.

Reference has also been made to the assumption, often necessary in practice, regarding the limits of the class of admissible hypotheses.

https://doi.org/10.1017/S030500410001152X Published online by Cambridge University Press