STATISTICAL THEORY OF PROPHYLACTIC AND THERAPEUTIC TRIALS

I. LIMITATIONS OF THE UNIQUE NULL HYPOTHESIS

BY

LANCELOT HOGBEN AND RAYMOND WRIGHTON

Department of Medical Statistics, University of Birmingham

1. PRELIMINARY CONSIDERATIONS

Until a comparatively recent date, available literature on the efficacy of therapeutic procedures exclusively recorded the outcome of a *follow-up* at consultant level, based as such on case histories of patients willing to testify or report for examination. Commonly with no control group for comparison, commonly also with scant regard for biased selection of cases treated, and rarely undertaken with precautions to validate the testimony of the patient or the clinical judgement of the author, a clinical trial so conceived violates any or all of three canons of scientific method, as is now becoming recognized widely in all branches of medicine, except perhaps psychiatry. A change of outlook is largely due to the impact of more exacting standards of evidence established at an earlier date in connexion with the assessment of prophylactic measures, partly as a consequence of public controversy over the merits of vaccination.

In this context, the term prophylactic calls for no comment. We here employ the expression *therapeutic* measures in the widest sense, including administration of drugs or convalescent sera, operative and manipulative surgery, diathermy and radiation (deep X-ray, short wave, radium) treatments, occupational and physiotherapy (including remedial gymnastics, faradization, massage), rehabilitation techniques. Bradford Hill (1951), who has himself directed a series of therapeutic trials on the now familiar pattern expounded by Greenwood (1935), has lately set forth in clear and simple language some of the essential safeguards of a scientific assessment of remedial measures; and there is no need to recapitulate them in this context. Our aim in what follows is to examine statistical procedures invoked to validate results within a framework of the precautions to which he has drawn attention; but it will simplify our task if we first specify the desiderata. Prophylactic or therapeutic trials alike pre-suppose:

- (i) a prescription of treatments;
- (ii) a yardstick of comparison, *i.e.* a control group untreated or treated by a method alternative to the test method;
- (iii) a criterion or criteria of efficacy;
- (iv) a method of scoring (iii);
- (v) an agreed basis for selection of test subjects;
- (vi) a statistical procedure for validation of (ii), (iv), and the conclusions suggested by the method of scoring employed.

The treatment prescription is worthy of comment only in so far as it sets a limit to the legitimacy of any positive assertion we may make. Thus it may specify different drugs, different dosages of the same drug, or the same drug with different ancillaries (e.g. differential regulation of fluid intake). If we merely specify *Treatment A* and *Treatment B* in terms of two different drugs, the outcome of our trial can merely tell us, if anything, whether one drug is preferable to the other within the framework of the prescribed dosage. This is not a statistical issue, and need not detain us. With respect to the second, there is little to add to the canons which Bradford Hill has propounded. The third calls for comment because the choice of the criterion and the assessment of its value may raise issues of statistical as well as of clinical interest. Briefly, we may classify our criteria of efficacy as follows:

- (1) Attack rate (*prophylaxis only*)
- (2) Severity:
 - (a) mortality
 - (b) complications
 - (c) sequelae
 - (d) particular signs and/or symptoms
- (3) Laboratory (*biochemical, pathological, radiological, haematological, etc.*) or other objective (*height, weight, body temperature*) tests
- (4) Clinical judgement
- (5) Patient's subsequent testimony
- (6) Duration of treatment, signs or symptoms
- (7) Relapse rate

Of the foregoing, (3) raises *ad hoc* issues of statistical analysis mentioned at a later stage. Both (4) and (5) will be more or less valuable or worthless in so far as the design of the trial takes within its scope provision for statistical assessment of their reliability. On this topic, there is again little to add to what Bradford Hill has written about it. It is pertinent to remark that:

- (a) duration of stay in hospital has special pitfalls, for reasons adequately set forth by Mackay (1951);
- (b) the reason for putting (7) last on our list is a reminder that recourse to only one criterion of efficacy may lead to a distorted view of the merits of a new treatment.

The method of scoring is of more direct consequence for what follows. Among ways in which we may express the outcome of a trial in numerical terms amenable to statistical treatment, it is customary to distinguish two main types, subsumed by the terms statistics of attributes and statistics of measurement, or, more briefly, qualitative and quantitative statistics. Each form of words is misleading and we shall here adopt the following dichotomy:

- (i) *Taxonomic scoring*, when we specify a group numerically by the number(s) of individuals assignable to a class (or classes);
- (ii) *Representative scoring*, when we specify a group numerically by some average of a range of score values individually assignable to its members.

Under (i) our criterion of *class* membership may be qualitative or quantitative. Under (ii) the score assigned to each individual of the group may be a count or a measurement. Neither the dichotomy qualitative-quantitative nor the dichotomy

attribute-measurement therefore conveys clearly the essential distinction which the following example illustrates. If we score a group of patients by the number or proportion of individuals with an *r-b-c* of over 4.10^6 per c.mm., our method of scoring is taxonomic, but our class criterion is essentially quantitative. If we score the same group by the mean number of red cells per c.mm., our method of scoring will be representative, but the incorporated scores assigned to the individual will not be measurements *sensu stricto*.

To a large extent the criterion of efficacy will determine the choice of the method of scoring. Thus (1), (2), (3), and (7) above necessarily invoke the taxonomic method; but either method is applicable to (6) and in some measure to (3), e.g. antitoxin titre in prophylactic trials. Where choice is open, as the foregoing remarks on the *r-b-c* illustrate, the issue is not primarily statistical; and it is not trivial to say so, since the statistician is all too prone to assume the contrary. In some situations, the representative method may have much to commend it in terms of available statistical techniques. The statistician, off his guard, may therefore be likely to prefer it as more *sensitive* in his own jargon; but the enumeration of individuals will be commonly more consistent with the humane end in view than will a method of scoring which records the average of the treatment group as a whole.

As stated, a satisfactory basis for selection of personnel raises issues with the claims of which the consultant clinician, accustomed to contact with a highly selected population of sick persons and with few opportunities for examining a control group, has only lately come to terms. For the most part, these are commonplace to the laboratory worker trained in the discipline of controlled experiment; but there is still need to emphasize an antinomy inherent in the practice of the laboratory, and statistical principles sometimes invoked unjustifiably in theoretical analysis of experimental data.

When we compare two treatments we may proceed in two ways. One is to choose individuals for each treatment group at random on the explicit assumption that each such group is a representative sample of a *homogeneous universe**. The other is to pair off individuals sharing common characteristics (e.g. age, sex, body build), or successive observations referable to one and the same individual, on the implicit assumption that each individual or observation referable to one and the same treatment group is a unit sample from a stratified universe. If we anticipate that the outcome of such pairing will be a contrast so clear-cut as to dispense with the need for statistical validification, the design of a laboratory experiment or of a field trial in conformity with the method of stratified sampling has everything to commend it; and the inclinations of the experimentalist will commonly favour the procedure.

* It is important to realize that we here speak of a universe as homogeneous, or contrariwise as stratified, in the statistical connotation of the terms. If we classify a population by age groups and take samples of predetermined size from each, we may legitimately speak of our sampling system and/or universe as stratified in a descriptive sense consistent with the possibility that the expected sample score is the same for each age group (sub-universe or stratum). If each stratum of the universe is indefinitely large and identical with every other one with respect to relevant parameters, the effect of successively combining unit samples from *r* such strata is statistically equivalent to that of successively taking an *r*-fold sample from any one of them. From the statistical viewpoint we then regard the universe as homogeneous.

What too few laboratory or field workers sufficiently realize is that:

- (a) the algebraic implications of stratified sampling are often intractable;
- (b) many prescribed statistical techniques are valid, only if we postulate a homogeneous universe as the source of each sample available for comparison.

In a later contribution in this series, we may have occasion to explore this aspect of the design of a trial more fully; and to clarify the proper choice of a technique of validification with special reference to circumstances in which pairing is or is not consistent with its essential postulates. In any case, we have still to consider the rival claims of alternative methods of validification. These we may provisionally distinguish as:

- (a) *decision tests*, ostensibly devised to adjudicate upon the merits of particular hypotheses;
- (b) *techniques of estimation*, the aim of which is to make legitimate statements about a parameter (or parameters) of a universe or sub-universe on the basis of information contained in a sample.

At this stage, we intentionally sharpen a distinction which we shall refine in a later communication. Here it suffices to say that we might alternatively define (b) as setting limits to what hypotheses are admissible on the basis of a set of observations. The choice of the word *merits* in preference to *truth* is also deliberate, inasmuch as it is necessary to make a rough and ready distinction* between different targets of statistical reasoning in the domain of test decisions:

- (i) to decide whether to regard a particular hypothesis as correct or false;
- (ii) to limit the risk incurred by rejecting it, if it is indeed correct.

As we shall subsequently make more explicit, the limitation of risks incurred by rejecting either of a particular pair of hypotheses if true gives no assurance of a high probability of decision in favour of a third one which happens to be correct. It is therefore useful to make a broad distinction between statistical inference as conditional or unconditional. With this end in view it is here fitting to make the meaning of the term statistical inference as explicit as need be. In the most exacting sense of the term, and as the writers would prefer to use it exclusively, statistical inference has as its end in view† an unequivocal assertion coupled with a numerical specification of the long-run frequency of its truth within an assumed framework of indefinitely protracted repetition. We may speak of this numerical specification as the *uncertainty safeguard*, and may then distinguish the two types of statistical inference mentioned above as follows:

- (I) *unconditional*, if the uncertainty safeguard unconditionally specifies the probability of the falsity of the assertion;
- (II) *conditional*, if the uncertainty safeguard specifies the probability of the falsity of

* Anscombe (1951) makes a three-fold distinction:

It is worthwhile to distinguish different purposes one may have in accepting a hypothesis: (i) to base an administrative decision on, (ii) for further testing and confirmation, (iii) for acceptance into the corpus of scientific knowledge, to be relied on in future work. There are risks, variously assessable, in coming to decisions in all three cases. For example, in case (iii), if the hypothesis is later found to be seriously false a lot of effort in investigating other points may have been wasted. Just as with prior confidences, risks are rather vague in magnitude, but in a formal theory it would be tempting to postulate a complete numerical risk-function.

† Some contemporary writers use it in a more extended sense, embracing situations in which the conditional framework of (ii) involves a specification referable to the data themselves, and definitions of probability referable to introspective judgments unrelated to the long-term frequency of correct assertion.

the assertion within a framework of explicitly stated conditions independent of the data.

In symbolic form we may express the unconditional probability of false assertion as $P_f = (1 - P_t)$, and the conditional probability of false assertion within the framework of a particular Hypothesis A as $P_{f.a} = (1 - P_{t.a})$. Thus an assertion of the form $P_t = x$ is an example of unconditional statistical inference. Besides these two forms of statement we may make one of the form $P_t \geq x$. At the epistemic level of communication (Meredith, 1951), the more exact assertion, $P_t = 0.95$, has no priority over the less definite assertion, $P_t \geq 0.95$; and we may prefer to regard a statement expressed in the form $P_t \geq x$ as an example of (I), if we deem $(1 - x)$ to be an *acceptable* level of uncertainty. On the other hand, it serves no useful purpose to make an assertion of the form $P_t \geq 0.30$ if we regard any figure above 5 per cent. as an unacceptable level of uncertainty. Hence we shall have no practical interest in stating an inference of the form $P_t \geq x$ unless we should be content with the assertion $P_t = x$. Otherwise, any useful statement of statistical inference we may undertake conforms to (II).

In a subsequent communication we shall draw attention to neglected possibilities of the use of estimation procedures for the validification of evidence supplied by prophylactic or therapeutic trials. In this one our aim is to dispel widely current misunderstanding concerning what conclusions we may legitimately draw from the outcome of statistical tests in general use. Accordingly, we shall now examine the impact on statistical theory of considerations traceable to the practice of quality control in commerce and manufacture. Until the works of Wald (1947) and Neyman (1950) were published, the new concepts with which we shall here concern ourselves were only accessible in such highly specialized publications as those of Neyman (1937), Wald (1942), and Eisenhart and others (1947). Even among mathematical statisticians, probably only a few have as yet fully digested how radical a revision of widely current dogmas concerning the credentials of test procedures the new concepts must inevitably provoke when there is wider recognition of their logical implications. Since their implications are indeed of the utmost importance to a proper evaluation of universally current test procedures in the domain of experiment, we shall not hesitate in what follows to make use of numerical examples and simple models to get their true meaning into focus.

While we do not deny the legitimacy of a different approach to our theme in the context of the research laboratories of a commercial firm concerned with the manufacture of drugs (*vide* p. 114 *infra*), we here conceive our end in view as making an unequivocal statement with reference to the relative merits of *Treatment A* and of *Treatment B*. Inescapably, some sort of uncertainty safeguard should accompany such a statement; and the attitude we adopt to the theory of probability will dictate its form. In any case, we here assume that the uncertainty safeguard is expressible as the probability that our statement is false; and that the measure of this probability is referable to a limiting ratio realizable in an infinitely protracted series of trials carried out within the same framework. In previous remarks we have drawn a

clear-cut distinction between conditional and unconditional inference. More precisely, four variants of any safeguard conceived in statistical terms, are open for discussion. We may be able to cite:

- (a) *without restriction* the probability that our assertion is false.
- (b) without *other* restriction that the probability of false statement so defined does not exceed an upper limit.
- (c) separately probabilities of false assertion each conditional on one of a *comprehensive* set of prior hypotheses.
- (d) separately also probabilities of false assertion each conditional on one of a set which is *not* comprehensive.

2. HEURISTIC VALUE OF THE DUAL HYPOTHESIS

For the past generation research workers engaged on agricultural trials, tests of the efficacy of therapeutic or prophylactic measures, sociological field work, and bio-assay have relied for validation of their results on tests devised by Fisher (1925) and his co-workers in conformity with a familiar pattern. Such tests entail:

- (a) the invocation of a unique so-called *null* hypothesis which prescribes the frequency with which a sample score will lie outside a prescribed limit (or limits);
- (b) the specification* of a criterion of rejection, *i.e.* the convention to reject the hypothesis if the sample score does in fact lie outside the prescribed limit(s).

Customarily (and oddly) the corresponding limiting frequency adopted is at the 95 per cent. (approximately 20:1 odds) level; and the possibility of defining it in such terms resides in the fact that the unique hypothesis chosen for the purpose has an assignable distribution function.

From one viewpoint, the prevalence of the fashion referred to is understandable. The text of Fisher (1925) prepared the way for manuals by Snedecor, Tippett, Hagood, Quenouille, and others, exhibiting schemata for computation in conformity with the Fisher test prescriptions. By recourse to a wealth of exemplary material the research worker willing to take the test prescription on trust can therefore readily, it may be all too readily, select a type of specimen at least seemingly like his or her own problem. None the less, there must be among those who do so, not a few who have felt misgiving for any (or all) of several reasons, notably the following:

- (a) not infrequently the form of the null hypothesis is irrelevant to the main issue, as when the decision that two treatment procedures have different results is of trivial interest in comparison with the decision that *Treatment B* is at least *so much* more effective than *Treatment A*;
- (b) the type of decision which concerns the investigator determines the choice of a particular null hypothesis far less than considerations of algebraic convenience *vis à vis* the specification of a sample distribution;
- (c) the test prescription takes no stock of any alternative hypothesis which may indeed be the main concern of the investigator, as when the disadvantages of

* In a later communication we shall have occasion to emphasize a difference between the opposing views of test procedure advanced by Fisher and his adherents on the one hand, and by the school of Neyman, Pearson, and Wald on the other. In the terminology of Meredith (1951), their concepts are referable to different epistemic levels, Fisher's to weighing evidence (already obtained), that of the others to prescribing rules of inference in conformity with a prescription independent of the observations. For this reason, Fisher's so-called exact 2×2 test lies outside the scope of the concepts discussed in Section 3 below.

dismissing a new treatment if more efficacious outweigh the propriety of cautious adherence to an established procedure.

The first misgiving has special reference to the domain of estimation, and as such to the theory of interval estimation specially associated with the name of Neyman (1937). Its relevance to the conduct of prophylactic and therapeutic trials will be the theme of a subsequent communication. The second and third raise issues which a theory of test procedure also advanced by Neyman and Pearson has brought into focus; but up to the present their critique of the unique null hypothesis has exerted little influence on research workers outside America. This is less because their writings lack the polemic vitality of their predecessors than because the concepts invoked are logically subtle and on that account difficult to assimilate unless examined against a background of familiar material. What follows claims to little originality, the aim being to help the laboratory or the field research worker to recognize pitfalls in previously accepted test procedures and to materialize some of the essentially novel concepts of the Neyman-Pearson approach.

One way of doing this is to formulate a biological problem which involves no predilection for a single hypothesis such as the *null* one in virtue of algebraic convenience as such. Accordingly, we shall think of a culture of *Drosophila* containing normal females and females which carry a sex-linked lethal gene. With Bacon we shall concede that nature is more diverse in her operations than man in his conceptions, but our knowledge of laboratory conditions (presumptively highly standardized) will justify the provisional assumption that any such female fruit-fly with an excessively large number of female offspring will in fact be either an entirely normal female or a lethal carrier. That is to say, we exclude such a contingency as the possibility that there is an endemic rare virus disease more fatal to male than to female larvae. We may then with justifiable assurance postulate two hypotheses about any female in the culture:

Hypothesis A: The female is normal, in which event the probability that any one of her offspring will be female is $p_a = \frac{1}{2}$.

Hypothesis B: The female is a lethal carrier, whence the probability that any one of her offspring will be female is $p_b = \frac{2}{3}$.

We shall now suppose that a particular female has 144 offspring, and examine the current theory of test procedure when the end in view is to decide whether we shall adopt one or other hypothesis. Our initial concern will thus be with what the test prescribes, and as such has no necessary connexion with whether it leads us to a correct decision. We first note that each hypothesis equally prescribes for 144-fold fraternities referable to a single fly mother the long-run frequency of such as respectively contain 0, 1, 2, 143, 144 females. We may in fact specify the relevant parameters thus:

Hypothesis	Size of Sample (fraternity)	Probability that any Single Offspring is Female	Mean No. of Females in Sample Fraternity	S.D. of Score Distribution of Sample
A	144	$p_a = \frac{1}{2}$	$M_a = 72$	$\sigma_a = 6.00$
B	144	$p_b = \frac{2}{3}$	$M_b = 96$	$\sigma_b = 5.66$

From an algebraic viewpoint neither hypothesis specified above has anything to commend it as preferable to the alternative; but we may lazily and arbitrarily agree to consider first of all the consequences of adopting Hypothesis A as the *null hypothesis* in the traditional sense, if only because laboratory and field workers would commonly do so in a comparable situation. Lazily and arbitrarily also, we shall first adopt a *modular* criterion of rejection or acceptance for the same reason, *i.e.* we shall reject the hypothesis chosen unless the number of females x is such that:

$$|(x - M_a)| \leq X_a$$

In conformity likewise with current convention, we shall choose the score X_a so that the probability (α) that x will lie in the *critical region*, *i.e.* outside the range specified above, is about 0.05, *if the null hypothesis correctly describes the situation*. For samples of 144 and values of p_a (or p_b) anywhere (as in this example) within the range 0.1 to 0.9, the normal integral gives an adequate quadrature at the so-called significance level 95 per cent., if we make the appropriate half-interval correction. If we choose $X_a = +12.5$, so that $(x - M_a) = \pm X_a$ when $(x - M_a) \simeq \pm 2.08\sigma$ the table of the normal integral sets $\alpha \simeq 0.038$. We have then made the decision to regard a female with 144 offspring as normal if the number of her female offspring lies in the range of 60 to 84 inclusive and to reject her claims as such, *i.e.* in this context to regard her as a lethal carrier, if her female offspring number more than 84 or less than 60.

In the Neyman-Pearson theory of test procedure we speak of α as the conditional probability of making an error of the *first kind*. Now the cited value of α ($\simeq 0.038$) correctly assigns the probability of rejecting the null hypothesis only on the assumption that the latter is true, *i.e.* that the mother fly is normal. This we do not know, the aim of the test being to throw light on the alternative possibility. If we carry out the rule of the test consistently, we shall sometimes make an error of the first kind, *i.e.* reject normal flies as such and by the same token wrongly identify as carriers flies which are indeed normal. Conversely, we shall sometimes apply the test to flies which are indeed carriers. If the number of females among their progeny lies within the range 60–84 inclusive we shall reject them as such. We shall then wrongly accept the null hypothesis. This is the error of the second kind, which we make in this context if the relevant parameters of the appropriate distribution are:

$$p_b = \frac{2}{3}, M_b = 96, \text{ and } \sigma_b \simeq 5.66.$$

With due regard to the half-interval correction, the region we then exclude is from 59.5 to 84.5, bounded by $(x - M_b) = -36.5$ and $(x - M_b) = -11.5$, *i.e.* $(x - M_b) \simeq -6.4\sigma_b$ and $(x - M_b) \simeq -2.03\sigma_b$. Since the area of the normal integral of unit variance from $-\infty$ up to -6.4 is utterly trivial, we make no sensible error if we say that the consistent application of the rule leads us now to reject carriers as such with a probability (β) assigned by the area of the the normal curve of unit variance in the range from $-\infty$ to -2.03 . One speaks of this loosely as the probability of making an error of the second kind; and the table of the normal integral in this case cites the value $\beta \simeq 0.0212$. More explicitly, β is the *conditional* probability

of accepting the null hypothesis, *if* it is false; but we have as yet said nothing about how often it will be.

In short, the only information we have at our disposal so far bears on the probability (α) of rejecting the null hypothesis when it is true, and that (β) of rejecting the alternative when the latter is true, *i.e.* of accepting the null hypothesis *when* it is false. If we now suppose that we actually know the proportion of normal and carrier females in the culture, we can take our analysis a decisive step forward. We shall assume that the culture consists of 500 mothers, of which 450 are normal and 50 are carriers. If we then choose at random* any single fly with 144 offspring as a test subject we may say that:

- (i) $P_a = 0.9$ is the probability that it will be normal, *i.e.* the probability that the null hypothesis is applicable to the test subject;
- (ii) $P_b = 0.1 = (1 - P_a)$ is the probability that it will be a carrier, *i.e.* that the alternative hypothesis is applicable to the test subject.

We now have all the relevant data for a statistical specification of the long-run frequency of all four possible results of the outcome of the test:

- (1) The fly is normal and we rightly accept it as such:

$$P_a (1 - \alpha) = (0.9) (0.962) = 0.866$$

- (2) The fly is normal and we wrongly reject it as such:

$$P_a \cdot \alpha = (0.9) (0.038) = 0.034$$

- (3) The fly is a carrier and we rightly accept it as such:

$$P_b (1 - \beta) = (1 - P_a) (1 - \beta) = (0.1) (0.979) = 0.098$$

- (4) The fly is a carrier and we wrongly reject it as such:

$$P_b \cdot \beta = (1 - P_a) \beta = (0.1) (0.021) = 0.002$$

To each assertion consistent application of the rule leads us to make we may thus assign a probability that it will be true or false. We may then set out a balance sheet as follows:

	Assertion True	Assertion False
Null Hypothesis True	$P_a (1 - \alpha) = 0.866$	$P_a \cdot \alpha = 0.034$
Null Hypothesis False	$(1 - P_a) (1 - \beta) = 0.098$	$(1 - P_a) \beta = 0.002$
Total	$1 - \beta - (\alpha - \beta) P_a = 0.964$	$\beta + (\alpha - \beta) P_a = 0.036$

In conformity with the definition given above, we may legitimately speak of the probability (P_t) of making a correct decision, or the probability (P_f) of making a false one, by consistent application of the rule, in which case our balance sheet yields:

$$P_t = 1 - \beta - (\alpha - \beta) P_a = 96.4 \text{ per cent.} \quad \dots (i)$$

$$P_f = \beta + (\alpha - \beta) P_a = 3.6 \text{ per cent.} \quad \dots (ii)$$

For reasons we shall come to later, the outcome of our choice of a rejection criterion is here vastly more encouraging than need be in most situations; but we can do better. We have lazily adopted a *modular* criterion because laboratory and field

* The effect of the lethal gene on the fertility of the fly introduces a bias for which we can allow, and one which we may therefore deliberately neglect for *heuristic* purposes.

workers commonly do so, regardless of the end in view, when the sample distribution prescribed by the null hypothesis is symmetrical. Now fraternities of 144 flies of which less than 60 are females will be vastly less common if the mother is a carrier than they would otherwise be. It would thus seem to be more reasonable to restrict our attention to families with an *excessive* number of females. We shall now therefore adopt a *vector* criterion, *i.e.* reject as normal only mothers with more than 84 female offspring, we exclude only one tail of the approximately normal distribution and halve our error of the first kind, *i.e.* set $\alpha = 0.019$. For reasons stated this does not materially affect the value of β , since the chance that a carrier will have less than 60 females among 144 offspring is negligible. If we then say that we shall reject the null hypothesis at the vector level $+2.08\sigma$ in contradistinction to the modular level $\pm 2.08\sigma$, we now put $\alpha = 0.019$ but $\beta = 0.021$ as before. Whence our balance sheet summarized by Equations (i) and (ii) becomes:

A. Modular Criterion. Reject the null hypothesis if:

B. Vector Criterion

FIG. 1.—Testing two hypotheses.
Null Hypothesis A. $P_a = 1$
Alternative B. $P_b = 0$

$$P_t = 0.981 = 98.1 \text{ per cent.}$$

$$P_f = 0.019 = 1.9 \text{ per cent.}$$

That the adoption of the vector criterion does in fact give a better prognosis of correct decision is not surprising; and Fig. 1 sufficiently exhibits why this is so in the situation under discussion.

At this stage we may also note with profit an interesting consequence of (i). If $\alpha = \beta$ so that $1 - \beta = 1 - \alpha$ and $(\alpha - \beta) = 0$, Equation (i) reduces to $P_t = (1 - \alpha)$. Within the framework of our assumptions that there is only one admissible alternative to the null hypothesis, so that $P_b = (1 - P_a)$, we can assign a value to the long-run frequency of correct decision based on consistent application of the rule without any prior knowledge (P_a or P_b) of the population at risk if we define our rejection criterion in such a way as to equalize the probabilities of errors of the two kinds. We can then predetermine that the value of α may be as

small as we care to make it by prescribing a *sample size sufficiently large*. Needless to say, this presupposes the possibility of defining the distribution function of the single admissible alternative hypothesis.

Within the same framework of assumptions and in the same model set-up, let us now explore the effects of making our criterion for rejecting the null hypothesis more exacting, in the sense that our error of the first kind is less. Thus we shall decide to accept a female fly with 144 offspring as normal if (*vector* criterion) she has 88 or less female offspring and deem her (rightly or wrongly) to be a lethal carrier if she has 89 or more. We then set the decision limits on either side of $x = 88.5$, in which event: $(x - M_a) = +2.75\sigma_a$, and $(x - M_b) = -1.33\sigma_b$.

Whence from the table of the normal integral we derive $\alpha = 0.003$ and $\beta = 0.092$. If we paint in these values in Equation (i) we get:

$$1 - \beta = 0.908 \text{ and } (\alpha - \beta) P_a = -0.080$$

$$\therefore P_t = 0.988 \text{ or } 98.8 \text{ per cent.}$$

In this situation little advantage (98.8 as against 98.1 per cent.) accrues from making our criterion for rejection of the null hypothesis more exacting; but we have chosen our null hypothesis as the hypothesis with greater *prior probability* since the culture contains a great excess of normal flies. Let us then reverse the situation by postulating that the culture contains 450 lethal carriers and 50 normal among 500 female flies in all, i.e. $P_a = 0.1$ and $P_b = 0.9$. In this case $(\alpha - \beta) P_a = -0.009$, so that

$$P_t = 0.908 + 0.009 = 0.917 \text{ or } 91.7 \text{ per cent.}$$

If the null hypothesis is referable to the *smaller* population at risk (i.e. if it has lower prior probability than the alternative) the effect of making the rejection criterion more exacting is thus to *lower the probability of arriving at a correct decision*.

Before discussing how far this rule is of general application within the framework of our model assumptions, let us take stock of another highly relevant variable, viz. sample size. For a fixed size of sample the foregoing results have sufficiently emphasized what a visual diagram suffices to demonstrate (Figs 2-4), i.e. we cannot decrease the conditional probability (α) of an error of the first kind without increasing the conditional probability (β) of an error of the second kind and *vice versa*. It is also of importance to appreciate that we can make β for a pre-assigned value of α as small as we wish to make it, if we make the size of the sample large enough. Conversely, we can keep α at a pre-assigned level for a smaller sample only by making β larger.

Consider for example the consequence of applying the foregoing test to fraternities of 100, so that $M_a = 50$, $\sigma_a = 5$, $M_b = 66.6$, and $\sigma_b = 4.71$. If we make the vector criterion of acceptance or rejection conveniently near the 2σ level, we may set it on either side of a score level 60.5 in which event $(x - M_a)$ is approximately $2.1\sigma_a$ and $\alpha \simeq 0.036$. If so, $(x - M_b)$ is approximately $-1.32\sigma_b$ and $\beta \simeq 0.0934$. The value of α here agrees as closely as need be for exemplary purposes with the value chosen ($\alpha = 0.038$) for the 144-fold sample when $\beta \simeq 0.021$. Thus the effect of reducing the sample size is to increase more than 4-fold the probability of an error of the second for a corresponding probability of error of the first kind.

In this case we can make our two error risks nearly equal by setting our limits of

FIG. 2.—Testing alternative hypotheses.
Hypothesis A. Mean (M) = 19.8, σ_m = 5
Hypothesis B. Mean (M) = 36.2, σ_m = 5

rejection and acceptance for the null hypothesis on either side of the score $x = 58.5$; in which event the null hypothesis sets the upper limit of acceptance at $+1.7\sigma_a$ and the alternative sets the lower limit for rejection at $-1.74\sigma_b$. Thus $\alpha \approx 0.045$ and $\beta \approx 0.041$. If $P_a = 0.9$, as in our first example, $P_t \approx 95.9$ per cent. For the 144-fold sample we have $P_t \approx 98.1$ per cent. when the two conditional risks are nearly equal.

Before we take the next one, we may well retrace our steps.

FIG. 3.—Testing alternative hypotheses.
Hypothesis A. M = 19.8, σ_m = 5
Hypothesis B. M = 36.2, σ_m = 5

We made the arbitrary decision to designate as our null hypothesis the assertion that the mother fly is normal. Actually, we have given no reason for doing so; and we may pause at this stage to dispose of a misconception widespread among those who carry out routine tests within the framework of the unique null hypothesis. There is prevalent a somewhat naïve view that we choose our null hypothesis as a safeguard against wishful thinking, and that we make accordingly our criterion of rejection as exacting

as need be. On such a view our criterion of rejection is at best a *disciplinary convention*; and as such has nothing to do with *unconditional* statistical inference. Also, one can justify it as such only if one chooses the null hypothesis on the understanding that one wishes to fall backwards in preserving one's rectitude, *i.e.* if the null hypothesis is actually the one the investigator has reasons for believing to be false. Evidently no recipe that the best Mrs. Beeton can prescribe will indeed meet one's requirements in all situations. If experiments on laboratory stocks have convinced the investigator that a new therapy is preferable to a current procedure, the enthusiastic research worker will not reasonably impose on the null hypothesis a criterion of rejection as exacting as that of the investigator undertaking experiments to test the credentials of extra-sensory

perception. In conformity with current procedure, he will nevertheless invoke a null hypothesis of the same type in either situation, and with the same convention (e.g. 0.05 significance level) of rejection, if accustomed to rely on current cookery book recipes. The reason is that the cookery book recipe will commonly prescribe as the appropriate null hypothesis the one which commends itself to the mathematician for reasons which have nothing to do with

FIG. 4.—Testing alternative hypotheses.
Hypothesis A. $M = 19.8, \sigma_m = 5$
Hypothesis B. $M = 36.2, \sigma_m = 5$

FIG. 5.—Alternative hypotheses involving nearly equal parameters.

Hypothesis A. $M = 19.8, \sigma_m = 5$
Hypothesis B. $M = 20.8, \sigma_m = 5$

the operational intention of the scientific worker.

In the model situation discussed hitherto, we have, in fact, sidestepped the temptation to choose our null hypothesis for this reason, since it would be equally easy to adopt as such the postulate that the fly mother is a lethal carrier. A rejection criterion identical in terms of the conditional risk of error of the first kind, as is indeed the most we can specify within the framework of a unique null hypothesis, would then lead us to results numerically inconsis-

tent with those we have so far explored. The reader may check this assertion by reversing the role of the two hypotheses in the foregoing examples.

Partly because of the size of the samples chosen, previous tests in our model situation have led to a high probability of correct decision arrived at in conformity with traditional procedure, *i.e.* within the framework of the unique null hypothesis. This will lead us to a totally wrong view of what we can rely on it to accomplish, if we fail to take stock of two background conditions plausibly invoked in the prescribed set-up, but rarely admissible in other situations:

- we concede only *one* admissible alternative to the null hypothesis;
- we have postulated a complete specification of the sampling distribution in terms of the alternative thereto.

It will be simpler, if we first examine the implications of (b). In all examples hitherto cited, we have found that $P_t > 0.5$, i.e. that more than 50 per cent. of our decisions will be right if we consistently follow the last prescription, in which event we shall be more often right than wrong. Now there is no reason why this should be true of our situation, other than the fact that we can here fix in advance the size of the sample and the criterion of rejection or acceptance for the null one with due regard to the value of the relevant parameters of *both* hypotheses. To see the relevance of the consideration last stated, let us now replace the female carriers of a sex-linked lethal gene by females with a virus infection to which their male offspring succumb somewhat more readily than their sister flies. We shall postulate a sex ratio of 11:9 in favour of females among the progeny of infected mothers. Our alternative hypothesis is now that $p_b = 0.55$.

On the new hypothesis, which we again provisionally assume to be the only admissible alternative to the null one ($p_a = \frac{1}{2}$), we have $M_b = 55$ and $\sigma_b \simeq 4.98$ for fraternities of 100. If we set our rejection criterion on either side of $x = 60.5$, we have as before

$$\begin{array}{lll} (x - M_a) = +2.1\sigma & \text{and} & (x - M_b) = +1.1\sigma, \\ \text{whence} & \alpha = 0.018 & \text{and} & \beta = 0.864. \\ \text{Thus} & (1 - \beta) = 0.136 & \text{and} & (\alpha - \beta) = -0.846; \\ \text{whence, from Equation (i),} & & & \end{array}$$

$$P_t \simeq 89.7 \text{ per cent. when } P_a = 0.9$$

$$P_t \simeq 22.0 \text{ per cent. when } P_a = 0.1$$

This example illustrates the important role of P_a and $P_b = (1 - P_a)$ which define* the *populations at risk under each hypothesis*. If $P_a = 0.1$, we have $P_t = 22$ per cent. and $P_f = 78$ per cent., i.e. consistent application of the rule will lead us to be wrong more often than right when the sample is as small as 100, but we can fix the size of the sample to ensure that $P_t > \frac{1}{2}$ only if we can assign at least a lower limit to P_a , and then only if we can assign a value to β . Now we cannot assign a value to β unless we can specify the appropriate parameter (in this case p_b) or parameters of the sample distribution of the admissible alternative hypothesis. In any case, it will rarely happen that we can as easily conceptualize the meaning we here confer on P_a , and still more rarely that we can equip it with a numerical value.

The considerations last stated do not exhaust limitations of current test procedure within the framework of the unique null hypothesis. We have so far assumed a single admissible alternative, and we can very rarely make such an assumption with propriety. Indeed, much statistical enquiry goes on against a background of an infinitude of admissible alternatives to any null hypothesis we may advance. In our model set-up we have already adumbrated this complication. So we may now with profit postulate three admissible hypotheses. We shall then get into focus a general conclusion concerning the utmost we can legitimately infer in the domain

* The parameters P_a , P_b above, and P_c in what follows, represent the *prior probabilities* of Bayes's theorem. The assumption of their equality when we do not know their values (Bayes's scholium) is indeed the vulgar error of neglecting the population at risk.

of unconditional assertion from the outcome of a decision test referable to a unique null hypothesis in the absence of background information concerning admissible alternatives thereto. We may anticipate it by reinterpreting Equation (ii) above.

When $\beta > \alpha$ in Equation (ii), $(\alpha - \beta)$ is negative and $P_f > \beta$, i.e. the unconditional uncertainty safeguard is *no greater than the greater conditional risk*.

When $\beta < \alpha$, $(\alpha - \beta)$ is positive and $P_f > \alpha$, i.e. the unconditional uncertainty safeguard is *no less than the smaller conditional risk*.

The conclusions last stated refer to a situation in which only two hypotheses are admissible. An examination of a model situation in which three hypotheses are admissible will exhibit it as a particular case of a general proposition of limits, expressible as follows, on the assumption that our concern is with the long-run proportion of the true assertions we make on the basis of a test procedure of the type under discussion:

- (1) The worst that can happen is that we shall exclusively encounter situations prescribed by the hypothesis associated with the greatest conditional risk (ρ_g) of erroneous rejection.
- (2) The best that can happen is that we shall exclusively encounter situations prescribed by the hypothesis associated with the smallest conditional risk (ρ_s) of erroneous rejection.
- (3) Accordingly, our uncertainty safeguard P_f will be within the limits ρ_s to ρ_g and this means that we can set an explicit upper limit to it, if knowledge of the range of admissible hypotheses permits us to assign a value to ρ_g .

Our new model will be that the following three hypotheses are admissible with reference to our fruitfly culture; and we therefore start by assuming that we can define the relevant parameters and sample distributions referable to each of them:

Hypothesis A: The female is normal $p_a = \frac{1}{2} = \frac{3}{6}$

Hypothesis B: The female transmits a virus infection $p_b = \frac{1}{2} = \frac{3}{6}$

Hypothesis C: The female carries a sex-linked lethal gene $p_c = \frac{2}{3} = \frac{4}{6}$

We shall designate the proportions of the three types of female fruitflies as respectively P_a, P_b, P_c , so that $(P_a + P_b + P_c) = 1$. In real life there is no reason to exclude the possibility that a lethal carrier could transmit the virus; but we assume for argument that such flies are 100 per cent. resistant to it. We are free to select any one of the above as our null hypothesis; and shall first assume that our null hypothesis is $p = p_a$. In that event we may choose a single rejection criterion $x > x_{bc}$, thereby fixing the conditional risk of rejecting the null hypothesis when true as α , that of rejecting Hypothesis B when true as β , and that of rejecting Hypothesis C when true as γ . The unconditional probability of an error of the first kind is αP_a . We accept the null hypothesis erroneously, if we either reject Hypothesis B when true or reject Hypothesis C when true. Whence the probability of wrong acceptance is $\beta P_b + \gamma P_c$, and the probability of making a false decision of one sort or the other is:

$$P_f = \alpha P_a + \beta P_b + \gamma P_c \quad \dots \dots (iii)$$

This is equivalent to:

$$P_t = (1 - P_f) = 1 - \alpha P_a - \beta P_b - \gamma P_c$$

If our null hypothesis is $p = p_a$, our rejection criterion must make $\gamma < \beta$ since p_b lies nearer to p_a than does p_c . If we define it so that $\alpha = \beta$ (whence $\alpha > \gamma$) we may write (iii) in the form:

$$\begin{aligned} P_t &= 1 - \alpha (P_a + P_b) - \gamma P_c \\ &= 1 - \alpha (1 - P_c) - \gamma P_c \\ \therefore P_t &= 1 - \alpha + (\alpha - \gamma) P_c. \end{aligned}$$

Since $\alpha > \gamma$, on the assumption that $\alpha = \beta$, $(\alpha - \gamma)$ is positive and

$$P_t \geq 1 - \alpha \quad \dots \dots (iv)$$

We have here arbitrarily chosen as our null hypothesis $p = p_a$. In principle, the procedure would be alike if we chose as our null hypothesis $p = p_c$. If we choose the second hypothesis, it will be different, because we must now adopt a modular rejection criterion, *i.e.* we reject Hypothesis B in favour of Hypothesis A if $x < x_{ab}$, and reject Hypothesis B in favour of Hypothesis C if $x > x_{bc}$. We shall then denote by β_a the probability of rejecting the null hypothesis when true in favour of Hypothesis A, and by β_c that of rejecting the null hypothesis when true in favour of Hypothesis C. The probability of erroneous rejection is then $P_b (\beta_a + \beta_c)$. The probability of erroneous acceptance is $(\alpha P_a + \gamma P_c)$, and $P_f = P_b (\beta_a + \beta_c) + \alpha P_a + \gamma P_c$. If we choose our two rejection criteria so that $\alpha = \beta_a$, and $\gamma = \beta_c$, we then obtain

$$\begin{aligned} P_f &= (P_a + P_b) \beta_a + (P_b + P_c) \beta_c \\ &= (1 - P_c) \beta_a + (1 - P_a) \beta_c. \\ \therefore P_f &\leq (\beta_a + \beta_c). \\ \therefore P_t &\geq 1 - (\beta_a + \beta_c). \quad \dots \dots (v) \end{aligned}$$

Since the conditional risk of an error of the first kind is $(\beta_a + \beta_c) = \beta$, the choice of Hypothesis B as our null hypothesis leads us to a result equivalent to Expression (iv):

$$i.e. \quad P_t \geq 1 - \beta$$

Thus we can always choose our rejection criterion or criteria to make the uncertainty safeguard no greater than the conditional risk of error of the first kind, when either of only two alternatives to the null hypothesis is admissible, each being also discretely *specifiable*. With the same reservation we can make the probability of correct decision as near to unity as we care by making the size of the sample appropriately large.

So far we have presumed a backstage view of the situation, *i.e.* that we can in fact specify precisely each admissible alternative to the null hypothesis; but it will rarely happen that we can in fact do so. If we cannot specify either admissible alternative, we must retrace our steps, recalling that the form of the uncertainty safeguard does not depend on the choice of a modular or vector criterion of rejection. If ρ_h is the conditional risk of rejecting Hypothesis H when true, we may write it for any three hypotheses with discrete parameters p_0, p_1, p_2 as

$$P_f = \sum_{h=0}^{h=2} P_h \cdot \rho_h \quad \dots \dots (vi)$$

In this expression,

$$\sum_{h=0}^{h=2} P_h = 1. \quad \dots \dots (vii)$$

We shall denote by ρ_g the greatest and by ρ_s the smallest values of ρ_h without assuming

to which hypothesis either is assignable. We can then express either of the remaining values as:

$$\rho_g - e_{h,g} \text{ and } \rho_s + e_{h,s}$$

By definition, each value of $e_{h,g}$ and $e_{h,s}$ in the above is either positive or zero. On this understanding we may write our uncertainty safeguard alternatively as:

$$\begin{aligned} \sum_{h=0}^{h=2} P_h (\rho_g - e_{h,g}) &= P_f = \sum_{h=0}^{h=2} P_h (\rho_s + e_{h,s}). \\ \therefore \rho_g \sum_{h=0}^{h=2} P_h - \sum_{h=0}^{h=2} P_h \cdot e_{h,g} &= P_f = \rho_s \sum_{h=0}^{h=2} P_h + \sum_{h=0}^{h=2} P_h \cdot e_{h,s}. \\ \therefore \rho_g - \sum_{h=0}^{h=2} P_h \cdot e_{h,g} &= P_f = \rho_s + \sum_{h=0}^{h=2} P_h \cdot e_{h,s}. \\ \therefore \rho_g &\geq P_f \geq \rho_s. \end{aligned}$$

We may now generalize the utmost we can legitimately infer from the performance of a decision test within the framework of a unique null hypothesis, as we have been accustomed to perform it, *i.e.* against the background of an unknown, if finite, number (m) of admissible alternatives to it, and with no precise specification of any one of them. As before, we shall denote by ρ_h the probability of rejecting the h th member of the m -fold set when it is right, and by ρ_0 the corresponding risk of rejecting when right the hypothesis arbitrarily chosen as the *null* one ($p = \rho_0$), so that

$$P_f = \sum_{h=0}^{h=m} P_h \cdot \rho_h.$$

Consider now the hypothesis for which the risk of rejection (ρ_g) when true is *greatest*, so that $e_{h,g}$ is positive if $\rho_h = (\rho_g - e_{h,g})$ and $e_{g,g} = 0$.

$$\begin{aligned} P_f &= \rho_g \sum_{h=0}^{h=m} P_h - \sum_{h=0}^{h=m} P_h \cdot e_{h,g}, \\ &= \rho_g - \sum_{h=0}^{h=m} P_h \cdot e_{h,g}. \\ \therefore P_f &\leq \rho_g \text{ and } P_t \geq 1 - \rho_g \end{aligned} \quad \dots \dots (viii)$$

If ρ_s is the *smallest* value of ρ_h so defined, we may write $\rho_h = (\rho_s + e_{h,s})$, in which $e_{h,s}$ is again positive and $e_{s,s} = 0$, so that:

$$\begin{aligned} P_f &= \rho_s \sum_{h=0}^{h=m} P_h + \sum_{h=0}^{h=m} P_h \cdot e_{h,s}. \\ \therefore P_f &\geq \rho_s \text{ and } P_t \leq 1 - \rho_s \end{aligned} \quad \dots \dots (ix)$$

The situation last discussed suffices to get into focus the most we can say about the outcome of a single hypothesis decision test within the domain of *unconditional* inference, when, as commonly, we can neither state how many alternatives to it are each admissible nor specify each in numerical terms. If we denote by ρ_s and ρ_g respectively the least value and the greatest value of the probability of denying, in

conformity with the rejection criterion chosen, any one of the set of hypotheses when true, the probability of a false verdict lies between limits definable as:

$$\rho_s \leq P_f \leq \rho_g \quad \dots (x)$$

Unless we can specify the parameters of each admissible alternative to the hypothesis chosen as the *null* one, we can say no more about ρ_g in Expression (x) than it is less than unity; but if we can specify each admissible alternative hypothesis precisely, we can choose our rejection criterion to maximize the probability (α) of rejecting the null hypothesis when true, so that $\alpha = \rho_g$, whence $P_f < \alpha$. By appropriate choice of sample size we can then make the probability of a correct verdict *on the null hypothesis* as near to unity as we like, without invoking any information with reference to its *prior probability* (Neyman and Pearson, 1933). We thus arrive at the following conclusion: *the decision test procedure may be informative in the domain of unconditional inference but if, and only if, we can precisely specify each of an exhaustive and exclusive set of hypotheses.*

This vital restriction calls for further comment. A parameter p_h , definitive of an admissible alternative to the null hypothesis ($p = p_0$), may be indefinitely close to p_0 itself. If so (Fig. 5), $p_h \simeq 1 - p_0$ for samples of finite size; and we can gain no appreciable advantage by making $p_h = p_0$. This consideration has an important bearing on the concept of *test power* touched on below. Here it is relevant because we can rarely be certain that no such hypothesis alternative to the one chosen as null is indeed admissible. This raises a question of pivotal importance in connexion with the foregoing exposition: in what situations can one postulate an exhaustive and exclusive set of admissible hypotheses which fulfil all the relevant conditions now stated?

An important class, in which the postulate of an exclusive and exhaustive set of admissible hypothesis is legitimate, arises in pathology when we can:

- (i) classify test subjects as healthy or sufferers from a particular disease;
- (ii) assign a probability on the basis of laboratory experience to the assertion that a single test will fail to identify them as one or the other.

For heuristic purposes a criterion for screening tuberculous patients cited by Neyman will serve as a type specimen*. On the basis of laboratory experience, we assume that a single x-ray film will:

- (a) fail to detect the disease in 40 per cent. cases when present;
- (b) give a positive result for 1 per cent. of healthy test subjects.

Clearly we need to make more than one film, if we aim at a high level of satisfactory diagnosis. We shall assume that we take five and adopt as our test criterion the rule that we deem the disease to be present if at least one film is classifiable as positive. Our first, which we may call the null hypothesis (Hypothesis A), will be that the disease is present. Our test criterion leads us to reject the hypothesis if all five films are negative. The relevant parameter (p_a) is the probability of failure, in this case 0.4. Hence the probability of rejecting the null hypothesis (*i.e.* wrongly classifying the

* The medical specialist will recognize some arbitrary assumptions in the argument here advanced for illustrative purposes alone.

test subject as healthy) is

$$\begin{aligned}\alpha &= (0.4)^5 \\ &= 0.01024\end{aligned}$$

The alternative hypothesis (Hypothesis B) is that the test subject is healthy. If so, the probability (p) of getting a negative result from one film is 0.99, and that of getting five negative results is $(0.99)^5$. We shall reject the alternative hypothesis if at least one film is positive, *i.e.*

$$\begin{aligned}\beta &= 1 - (0.99)^5 \\ &= 0.04901\end{aligned}$$

The reader will find it instructive to explore the outcome of different test criteria based on different sizes of sample (*i.e.* numbers of films per test subject) within the framework of the foregoing assumptions. In this context, P_a in Equation (i), on p. 97, is the incidence of tuberculosis in the population. Our truth equation is

$$\begin{aligned}P_t &= 1 - \beta - (\alpha - \beta) P_a, \\ &= 1 - 0.04901 - (0.01024 - 0.04901)P_a, \\ &= 0.95099 + (0.03877)P_a.\end{aligned}$$

Thus the test criterion ensures a probability of correct decision in a little more than 95.5 per cent. if the incidence of the disease is 1 per cent., and must inevitably ensure a figure more than 95 per cent. for an overall correct verdict. However, we shall now see that unconditional assertions may be of subsidiary interest in connexion with decisions of this sort, though the procedure illustrated may well have important applications in the domain of differential diagnosis.

3. DOMAIN OF CONDITIONAL INFERENCE

In the foregoing section we have examined a model situation, *viz.* a fruitfly culture, to throw light on the relevance of test procedure to unconditional inference, *i.e.* our concern has been to assign a probability to a correct decision for or against a hypothesis. On the assumption that the female deemed to be normal in this context is a new and valuable mutant, we might also formulate our problem in terms of conditional inference. Thus we may wish to curtail both the risk of letting lethal genes accumulate in our stock and the risk of destroying normal stock otherwise available for perpetuating it. Accordingly, we decide to screen our females by setting up a rejection criterion which will set an acceptable limit to the risk incurred in retaining a lethal carrier and an acceptable limit to the risk of losing an otherwise normal female which carries the mutant gene we seek to perpetuate.

We can likewise, and usefully, regard the issue at stake in a diagnostic test, such as the one Neyman cites, as comparable to decisions which arise in *quality control*, in so far as the theory concerns itself with hazards respectively, though not quite appropriately*, designated *producer's risk* and *consumer's risk*. The main

* *Producer's risk* is the risk of rejecting for sale consignments which do in fact attain the standard guaranteed (*Acceptable Quality Level*) by the manufacturer. *Consumer's risk* is the risk of releasing for sale consignments of a standard just at a level of inacceptability prescribed by the manufacturer. In a later communication, we hope to discuss the credentials of quality control techniques and concepts, and their relevance to the clinical trial.

preoccupation of the administration in the situation last discussed will in fact have less to do with an overall assessment of correct judgment than with the penalties of making mistakes of two sorts. To classify wrongly a tuberculous person is to deprive him or her of proper treatment. To classify wrongly a healthy person is likely to cause unjustifiable alarm and despondency. A test procedure which prescribes that neither risk exceeds what the authorities regard as admissible satisfies the practical demands of the situation from their viewpoint. We may state these demands explicitly in the form:

- (i) if the test subject is tuberculous, the risk of erroneous diagnosis must not exceed α ;
- (ii) if the test subject is healthy, the risk of erroneous diagnosis must not exceed β ;

Any unconditional statement we can legitimately make in this context presupposes the possibility of classifying the test subjects exclusively as of one or other type; but the administrative intention does not change, if we concede that an appreciable number of test subjects are unclassifiable by recourse to any available independent diagnostic criterion. The test need not then lead to consequences embarrassing to authorities content to disclaim responsibility for individuals unless definitely deemed to be healthy or tuberculous. Undoubtedly, there will arise in administration many comparable *costing* situations in which a conscientious claim for limiting the requirements of a test to such conditional assertions is admissible; but the propriety of such a procedure in the domain of scientific research is at least open to debate.

Recent literature on quality control techniques justifies the suspicion that some writers would advance the claims of conditional decision tests of this type as appropriate in the domain of the prophylactic or therapeutic trial. It is therefore pertinent to examine the relevance of the analogy between the end in view of the salesman and that of the research worker approaching a clinical trial against the background of laboratory experiments *in vitro* or on animals. In this situation the investigator will lightly incur neither the risk of losing credit for major discovery nor that of being discredited by subsequent enquiry. If content to follow the practice of the large-scale commercial corporation, he will therefore invoke a test procedure which will set appropriate limits to the risk of wrongly rejecting the alternatives:

- (i) his own assertion that *Treatment B* guarantees *b* per cent. more cures than *Treatment A* ;
- (ii) the assertion of an imaginary critic that *Treatment B* guarantees only *a* per cent. more cures than *Treatment A*.

By all too easy stages, statistical inspection then becomes a recipe for statistical careerism. The investigator and his putative opponent relinquish their proper relation as colleagues in the impersonal pursuit of truth to embrace a convention which safeguards the *amour propre* of each. The decision to make the best of a bad job in this sense involves an ethical issue which is not amenable to arguments likely to win universal assent; but it carries with it an implication which may well damp the enthusiasm of the convert. This will come into focus, if we here digress to clarify the Neyman-Pearson concept of *test power*.

In the taxonomic domain of the two-class universe we specify α and β in the following way for the r -fold sample, when the criterion for rejecting Hypothesis A ($p = p_a$), and hence for accepting Hypothesis B ($p = p_b > p_a$), is $x \geq t$:

$$\alpha = \sum_{x=t}^{x=r} r_{(x)} p_a^x (1 - p_a)^{r-x}; \quad 1 - \alpha = \sum_{x=0}^{x=t-1} r_{(x)} p_a^x (1 - p_a)^{r-x} \quad \dots \dots (i)$$

$$\beta = \sum_{x=0}^{x=t-1} r_{(x)} p_b^x (1 - p_b)^{r-x}; \quad 1 - \beta = \sum_{x=t}^{x=r} r_{(x)} p_b^x (1 - p_b)^{r-x} \quad \dots \dots (ii)$$

What Neyman calls the power function $F(p)$ of the test for the same size (r) of sample, and for the same criterion score (t) is picturable as the graph of the following function over the range $p = 0$ to $p = 1$:

$$F(p) = \sum_{x=t}^{x=r} r_{(x)} p^x (1 - p)^{r-x} \quad \dots \dots (iii)$$

It follows that:

$$F(p_a) = \alpha \quad \text{and} \quad F(p_b) = 1 - \beta \quad \dots \dots (iv)$$

For given values of r and t , the condition that $\alpha = \beta$ is, of course:

$$F(p_a) = 1 - F(p_b) \quad \dots \dots (v)$$

Having fixed any criterion for rejection of the null hypothesis (Hypothesis A), and having chosen the alternative hypothesis ($p = p_b$), we speak of $F(p_b)$ as the *power of the test*. This being $(1 - \beta)$ is the probability of rejecting the null hypothesis when it is false, on the assumption that *the alternative is the only admissible one*. One test prescription is more powerful than another if it has a higher power in this sense for a fixed value of α , i.e. if it assigns a lower probability to error of the second kind for the same probability of error of the first. If the two test prescriptions both invoke the same distributions, the test which employs a larger sample must be the more powerful one.

The reader will find it instructive to plot $F(p)$ against p for the following example of a test procedure. The null hypothesis is that $p = \frac{1}{2}$ when $r = 144$. The rejection criterion is $x > 82.5$. For the distribution prescribed by the null hypothesis the mean is 72 with $\sigma = 6$. Whence the criterion score in standard form is $(82.5 - 72) \div 6 = 1.75$. This excludes 4 per cent. of the area of the normal fitting curve, i.e. $\alpha = 0.04$. For this set-up we may tabulate as below:

TABLE I

p	M	$X - (82.5 - M)$	σ	$\frac{X - (82.5 - M)}{\sigma}$	β	$F(p) = (1 - \beta)$
1.0	60	22.5	5.9161	3.8032	>0.999	<0.001
0.9	66	16.5	5.9791	2.7596	0.996	0.004
0.8	78	4.5	5.9791	0.7526	0.773	0.227
0.7	84	-1.5	5.9161	-0.2535	0.401	0.599
0.6	90	-7.5	5.8095	-1.2910	0.099	0.901
0.5	96	-13.5	5.6568	-2.3865	0.013	0.987
0.4	102	-19.5	5.4544	-3.5750	<0.001	>0.999

The concept of test power is easily interpretable in the alternative domain of representative scoring. The simplest type of test is then one which invokes alternative hypotheses specifying the true mean score of each of two normal universes. Consider now the following model. We do not know the mean value (M) of the normal variate; but we do know that the variance of the score distribution of the parent universe is 2,500, whence it follows that (σ_m^2) of the mean of the r -fold sample is $2,500 \div r$. For the 100-fold sample $\sigma_m = 5$. Our null hypothesis is that $M = 18.2$. The standard score corresponding to a sample value (M_x) of the mean is therefore $(M_x - 18.2) \div 5$. To make $\alpha = 0.05$, we must make the deviation equal to $1.64\sigma_m$,

$$\begin{aligned} \text{i.e. } (M_x - 18.2) &= 8.2, \\ \text{whence } M_x &= 26.4. \end{aligned}$$

The alternative hypothesis, which makes $\beta = 0.05$, is that

$$(M_x - M) \div \sigma_m = -1.64,$$

so that $(26.4 - M) = -5(1.64)$. Whence the hypothesis is that $M = 26.4 + 8.2 = 34.6$. If our alternative hypothesis were that $M = 28.2$, the score deviation would be $(26.4 - 28.2) = -1.8$ or $-0.36\sigma_m$. At this level $\beta = 0.359$. To make the two risks equal when the sample size is 100 and the alternative hypothesis is $M = 28.2$, we must choose the sample value (M_x) definitive of our rejection criterion, so that:

$$\frac{M_x - 18.2}{5} = -\frac{(M_x - 28.2)}{5}.$$

In this case $M_x = 23.2$,

$$\text{i.e. } (M_x - 18.2) = 5 = \sigma_m,$$

so that $\alpha = 0.159 = \beta$. To equalize both risks as nearly as possible at the level $\alpha = 0.05 = \beta$, when the alternative hypothesis is that $M = 28.2$, we must enlarge our sample size (r) so that $\sigma_m = 50 \div \sqrt{r}$ in the identity:

$$\frac{23.2 - 18.2}{\sigma_m} = 1.64 \quad \text{---} \quad \frac{-(23.2 - 28.2)}{\sigma_m},$$

whence we get $\sqrt{r} = 16.4$, and $r = 269$ to the nearest integer.

We may generalize the rules of test prescriptions thus:

- (1) to fix α at the $h_a\sigma_m$ level, we make our test criterion

$$\frac{(t - M_a)}{\sigma_m} = h_a, \quad \text{so that } t = M_a + h_a\sigma_m \quad \dots\dots (vi)$$

- (2) to determine β in terms of $h_b\sigma_m$, we then have

$$h_b = \frac{-(t - M_b)}{\sigma_m} \quad \dots\dots (vii)$$

- (3) to equalize the two risks without changing the size of the sample, we make

$$\frac{t_o - M_a}{\sigma_m} = \frac{-(t_o - M_b)}{\sigma_m}, \quad \text{so that } t_o = \frac{(M_a + M_b)}{2} \quad \dots\dots (viii)$$

(4) for equal risks, we specify α in terms of $h_a\sigma_m$ by the relation

$$\frac{t_o - M_a}{\sigma_m} = h_a, \quad \text{so that } h_a = \frac{M_b - M_a}{2\sigma_m} \quad \dots\dots (ix)$$

(5) to equalize risks at the level $\alpha = 0.05 = \beta$ (so that $h_a = 1.64 = -h_b$), we must change the size of sample from r_1 to r_2 so that the new value of σ_m is $r_1^{\frac{1}{2}} \cdot r_2^{-\frac{1}{2}} \sigma_m$ and

$$1.64 = \frac{(M_b - M_a) \sqrt{r_2}}{2\sigma_m \sqrt{r_1}}, \text{ so that } r_2 = \frac{(3.28)^2 r_1 \sigma_m^2}{(M_b - M_a)^2} \quad \dots\dots (x)$$

The reader will easily adapt the foregoing to the situation in which we require two parameters to specify the normal universe of a hypothesis, viz. M_u and σ_u . For the simpler case under consideration, the power of the test $(1 - \beta)$ for any alternative to the null hypothesis is expressible in the form:

$$h = \frac{t - M_b}{\sigma_m} \quad \dots\dots (xi)$$

We may then write:

$$P(M_b) = \frac{1}{\sqrt{2\pi}} \int_h^\infty e^{-\frac{1}{2}c^2} dc \quad \dots\dots (xii)$$

We may now make an exploratory table for test design, on the assumption that $r = 100$, $\sigma_m = 5$, and $t = 26.4$, so that $\alpha = 0.05$ when $M_a = 18.2$ (Table II):

TABLE II

Value of M definitive of alternative hypothesis	Level of rejection (h) expressed as $h\sigma_m$	Corresponding value of β	Power of test criterion ($1 - \beta$)	Value of β when $\alpha = \beta$	Value of r when $\alpha = 0.05 = \beta$
18.4	-1.6	0.945	0.055	0.492	672,400
19.4	-1.4	0.919	0.081	0.452	18,678
22.4	-0.8	0.788	0.212	0.337	1,525
24.4	-0.4	0.655	0.345	0.268	700
26.4	0.0	0.500	0.500	0.206	400
28.4	0.4	0.345	0.655	0.154	259
30.4	0.80	0.212	0.788	0.111	181
32.4	1.20	0.115	0.885	0.078	133
34.4	1.60	0.055	0.945	0.053	102
36.4	2.0	0.023	0.977	0.034	81
38.4	2.4	0.008	0.992	0.022	66
40.4	2.8	0.005	0.997	0.013	55

Since the size of sample fixes the power of the test for a fixed value of α , we can set β at a predetermined level appropriate to any single chosen alternative of the null hypothesis only if we plot P_f for different values of r . Table III sufficiently illustrates the procedure; and the reader may find it instructive to check the arithmetic by recourse to the foregoing equations.

TABLE III

POWER FUNCTION ($P_f = 1 - \beta$) FOR THE SAME MODEL AS IN TABLE I TABULATED SEPARATELY FOR DIFFERENT VALUES OF SAMPLE SIZE r WITH THE SAME REJECTION CRITERION ($\alpha = 0.05$) FOR THE NULL HYPOTHESIS ($M = 18.2$) WHEN $\sigma_m = 5$ FOR THE 100-FOLD SAMPLE

At the head of the columns are score values (M_x) corresponding to the condition $\alpha = 0.05$ and values of σ_m for the appropriate value of r .

Hypothesis $M =$	Size of Sample			
	81	144	256	324
	$\sigma_m = 5.5$ $M_x = 27.3$	$\sigma_m = 4.16$ $M_x = 25.03$	$\sigma_m = 3.125$ $M_x = 23.33$	$\sigma_m = 2.6316$ $M_x = 22.52$
19	0.056	0.073	0.082	0.090
21	0.129	0.166	0.227	0.281
23	0.221	0.312	0.456	0.571
25	0.341	0.496	0.702	0.826
27	0.480	0.681	0.879	0.955
29	0.622	0.829	0.965	0.993
31	0.749	0.924	0.993	0.999
33	0.849	0.972	0.999	0.999
36	0.942	0.996	0.999	0.999
39	0.983	0.999	0.999	0.999
41	0.993	0.999	0.999	0.999

We are now in a position to see more clearly the implications of approaching the interpretation of the outcome of a prophylactic or therapeutic trial as one of accommodating the producer's risk and the consumer's risk in the theory of quality control. If we do so, we conceive the test procedure as a game of chance in which the investigator arranges the stakes to accommodate the inclinations of a wholly imaginary contestant. His assertion is that *Treatment B* guarantees b per cent. more cures than *Treatment A*, and this fixes the value of b . His fictitious opponent asserts that *Treatment B* guarantees only a per cent. cures*; but because his opponent is merely a figment of his own fears, all that he can say about a is that: $a < b$. If he conceives that his opponent is ready to deny any operational advantage ($a = 0$) for *Treatment B*, he may set his own risk as equal to that of his opponent at a much lower level for a fixed size of sample than will be possible if his opponent makes a far more moderate claim (e.g. $a = \frac{b}{4}$). Alternatively, the design of a test to equalize risks at one and the same level will prescribe the availability of larger samples, if he conceives that his opponent, having first denied any advantage, will subsequently concede that there is some.

Having chosen the form of his own assertion (here the numerical value of b), the only exclusive admissible alternative with which he can equip the imaginary disputant of his claim is $a < b$; but the alternative test procedure then prescribes recourse to an infinite sample as a prerequisite to a firm decision. Within the restricted framework of conditional inference, the alternative test procedure can thus offer no simple nor unique recipe for validifying the operational advantage claimed for a new procedure.

* To avoid periphrasis, we here use the word "cure" regardless of the criterion of efficacy.

In case the reader finds the foregoing argument obscure, it may help to clarify the issue, if we go back to the data of Tables II and III (in which $\sigma_m = 5$ when $r = 100$). We first suppose that:

- (i) Disputant A initially asserts that $M = 18.2$ and Disputant B initially asserts that $M = 26.4$;
- (ii) Both disputants initially agree to accept the outcome of a test which vindicates the claim of A if the 400-fold sample value of M_x exceeds 22.3 .

In that event each takes a 5 per cent. risk of being discredited. We next suppose that an arbitrator persuades Disputant A to concede that $M = 19$, and Disputant B to concede that $M = 25.6$, still taking equal risk on the outcome of a 400-fold trial. Disputant A will still lose his case if $M > 22.3$, but each disputant now incurs a 26.8 per cent. risk ($\alpha = 0.2676 = \beta$) of getting an adverse verdict.

To be sure, we may concede the existence of a situation in which there need be no ambiguity about the choice of the appropriate alternative. Thus we may suppose that costing considerations exclude the desirability of substituting *Treatment B* for *Treatment A* unless it guarantees at least a per cent. more cures. If the protagonist P of the superior claims of *Treatment B* makes the claim that b exceeds a by a specified amount, each assertion is definite, and a verdict is obtainable within the framework of conditional inference. However, this situation will arise only if protagonist P has the folly to claim more than suffices to achieve any immediate practical advantage from so doing.

Throughout this section we have assumed that the current theory of dual risk in quality control operates exclusively within the domain of *conditional* inference. Though Neyman and Pearson (1933) appear to entertain as contrary a view as possible, our own is clearly the implication of the exposition of the *Sequential Ratio Test* (Wald, 1947), since the latter must eventually lead to a decision in favour of one or other alternative hypothesis, each being false. In this connexion it is pertinent to cite the following remarks of the Columbia Research Group (1945):

... In experiments in pure science, in which risks of errors can hardly be fixed with an eye to practical consequences, the setting of risks must rest on such difficult concepts as the *a priori* weight of evidence in favour of the hypothesis in question.

It is unfortunate that other expositors of the Neyman-Pearson theory of test procedure are not equally explicit. Wald (1947), who acknowledges it as the parent of sequential analysis, confines himself to the conditional domain, as in the following statement (*italics inserted*):

For any given critical region W we shall denote the probability of an error of the first kind by α and an error of the second kind by β . The probabilities α and β have the following important practical interpretation ... In the long run the proportion of wrong statements will be α if H_0 is true, and β if H_1 is true.

A procedure interpreted in such terms is, of course, consistent with the entirely conditional administrative intention involved in the screening of a population with reference to tubercular infection. When this is indeed all we ask of a decision

test, we may certainly agree with Kendall (1946, p. 272):

The argument does not depend on the relative frequencies of occurrence of the hypotheses H_0 and H_1 . . . There is no concealed form of Bayes's postulate in this approach.

Unlike ourselves and the Columbia Research team cited above, Kendall himself does not appear to concede that the prior probabilities of Bayes's theorem are indeed relevant when our concern is with unconditional inference. Introductory remarks to the argument mentioned in the last citation, after emphasizing that $(1 - \alpha)$ merely defines the probability of accepting the null hypothesis when true, continue as follows (p. 270):

but what about the case when it is not true? We cannot ignore this case, for its possible existence is the very reason for carrying out the test.

To avoid overstatement, let us finally emphasize a distinction implicit in this discussion and explicitly recognized in the use of the epithet *pure* as applied to science in the foregoing comment of the Columbia Research Group. Whether our concern is with conditional or unconditional assertion depends less on the subject matter than on the end in view. If we invoke statistical procedure to screen statements worthy of assimilation in the corpus of scientific knowledge, the only assertions relevant to the intention are unconditional in the sense in which we here use the term. If we invoke it as a guide for day-to-day decisions in commerce or government, a conditional type of assertion is commonly appropriate to our requirements. Failure to appreciate the level at which the need for a clear-cut choice arises has been, and is, the source of perennial confusion. A still growing literature on statistical methods for bioassay bears witness to this confusion, which is resolved when we recognize that statistical preoccupations relevant to bioassay as an instrument of commercial production or legal inspection are essentially different from those which should carry weight in the conduct of physiological research with the advancement of knowledge as its goal. By the same token, the main preoccupation of a research staff attached to a pharmaceutical firm need not be the same as that of a university department, and the statistical tests appropriate to their different aims will then also be different.

4. A SIMPLIFIED MODEL OF CURRENT TEST PROCEDURE

A simple model will here clarify the implications of carrying out the significance test most widely practised in the context of the clinical trial, *viz.* the χ^2 (1 d.f.) test for the 2×2 table on the assumption that there is no treatment difference. If p_a is the true proportion of cures by *Treatment A*, and p_b that of cures by *Treatment B*, we may speak of k as the operational advantage of the second treatment if $p_b = p_a + k$. The conventional null hypothesis is that $k = 0$, and the current procedure is to fix the rejection criterion (most commonly so that $\alpha = 0.05$) with no concern for any alternative value that k may have. This restriction of the scope of the analysis is worthy of comment for two reasons:

- (1) the presumptive reason for performing any test is that available evidence points to a contrary conclusion, *viz.* $k > 0$;

- (2) the existence of a difference will be of academic interest alone unless $k \leq a_0$, the least operational advantage which would justify the *substitution of Treatment B for Treatment A in practice*.

As elsewhere, we may simplify our problem by taking a backstage view of the test procedure, *i.e.* we shall assume that we have a reliable figure p_a for our yardstick procedure, and that p_a has the particular value 50 per cent. Also for simplicity, we shall assume that the size (r) of each sample is the same. If $p_{a,s}$ and $p_{b,s}$ are our observed sample estimates of p_a and p_b , their differences being $d_s = (p_{b,s} - p_{a,s})$ $M_d = (p_b - p_a) = k$ is the expected value of d_s and the variance of its distribution is

$$\sigma_d^2 = \frac{p_a q_a}{r} + \frac{p_b q_b}{r} = \frac{2p_a q_a}{r} + \frac{k(1 - 2p_a - k)}{r}$$

When $p_a = \frac{1}{2}$, as we here assume, this becomes:

$$\sigma_d^2 = \frac{1 - 2k^2}{2r}$$

When $r > 20$, the normal curve will give a close fit for the distribution of d_s , on the assumption that p_a lies near $\frac{1}{2}$. If $p_a = \frac{1}{2}$, we may thus define a square normal standard score of unit variance (χ^2 for 1 d.f.) by the ratio

$$c_k^2 = \frac{(d_s - M_d)^2}{\sigma_d^2} = \frac{2r(d_s - k)^2}{1 - 2k^2}$$

When $k = 0$ we may write $c_k = c_0$. This is our null hypothesis (H_0). We shall consider the possibility that two alternatives are admissible,

H_1 , that $k = 0.10$ (10 per cent. advantage)

H_2 , that $k = 0.05$ (5 per cent. advantage), labelling c_k accordingly.

We then have:

$$c_0 = +d_s \sqrt{2r}; \quad c_1 = \frac{-(10d_s - 1)\sqrt{r}}{7}; \quad c_2 = \frac{-(20d_s - 1)\sqrt{r}}{\sqrt{199}}$$

Since our concern is with the possibility that $k > 0$, our rejection criterion for H_0 will be that $d_s > d_0$. If we adopt the convention $\alpha = 0.05$, we then put $c_0 = 1.64$, so that

$$d_0 = \frac{+1.64}{\sqrt{2r}}$$

When $r = 50$, we have then $d_0 = 0.164$. This being greater than either $M_d = 0.1$ (on H_1) and $M_d = 0.05$ (on H_2), $(d_s - k)$ is *positive* for each admissible alternative, so that

$$c_1 = \frac{0.64\sqrt{50}}{7} \simeq +0.65 \quad \text{and} \quad c_2 = \frac{3.28\sqrt{50}}{\sqrt{199}} \simeq +1.14.$$

For these values of the standard score, the table of the normal integral gives

$$\beta = 0.74 \text{ when } c = +0.65, \quad \text{and} \quad \beta = 0.87 \text{ when } c = +1.14$$

If our concern is with an overall verdict on the truth or falsehood of the null hypothesis, these figures speak for themselves in the light of what we have established in Section 2 above. If our concern is merely with the conditional risk of rejecting a good and of accepting a bad substitute for *Treatment A*, we can equalize the two risks by setting $c_0 = -c_1$ or $c_0 = -c_2$. For samples of 50, we then have:

- (1) $c_0 \simeq 0.5$ when $k = 0.1$, so that $\alpha \simeq 0.31 \simeq \beta$;
- (2) $c_0 \simeq 0.025$ when $k = 0.05$, so that $\alpha \simeq 0.40 \simeq \beta$.

In this way we can make a table of conditional risks for different values of r :

TABLE IV

r	$\simeq$ Value of β for $\alpha = 0.05$		$\simeq$ Value of $\alpha = \beta$	
	on H_1	on H_2	on H_1	on H_2
50	0.74	0.87	0.31	0.40
100	0.55	0.82	0.24	0.36
200	0.36	0.74	0.16	0.31
400	0.12	0.59	0.13	0.24

5. SUMMARY

(1) Hitherto it has been customary to assess the claims of therapeutic and prophylactic measures in statistical terms by recourse to tests which invoke a unique and so-called *null hypothesis*, namely that the procedures compared are equally efficacious.

(2) This procedure has no bearing on the operational intention of the trial, *viz.* to find out how much advantage accrues from substituting one treatment for another.

(3) Within its more restricted domain, the credentials of any significance test which takes within its scope only one hypothesis have now to meet the criticism that it takes into account only one sort of error, *viz.* that of rejecting the hypothesis when it is true.

(4) A procedure which justifies assertions of so limited and conditional a scope may be a useful self-disciplinary convention; but its claims to rank as an instrument of statistical inference are no longer acceptable.

(5) The Neyman-Pearson theory of *alternative* test procedure leaves open the possibility of prescribing:

(a) rules of *unconditional* statistical inference, as may be possible when we can specify the distribution function of every admissible hypothesis;

(b) rules of *conditional* statistical inference, when our legitimate concern is to safeguard ourselves against alternative hazards.

(6) Most commonly, situations in which decision test procedures are consistent with the restriction stated under 5(a) arise only when 5(b) is consistent with the practical aim of the test, *e.g.* when we approach problems of costing or screening from the viewpoint of the administrator; but the dual test conceived in terms of 5(a) may have useful applications in medicine, *e.g.* in connexion with differential diagnosis.

(7) In the context of the therapeutic or prophylactic trial, we can never restrict

the range of admissible hypotheses to accommodate the possibility of useful unconditional assertions. Both the propriety and the practicability of interpreting it *en rapport* with 5(b) are questionable, if the end in view is the advancement of science.

(8) The outcome of our critique is therefore a recognition of the need for a new approach to the validification of such trials by methods of *comparative estimation*. These will be the theme of a subsequent communication.

ADDENDUM. When this communication went to press, the authors were not aware of the views on test procedure expressed in two important contributions by Jackson (1936) and von Mises (1943). Jackson introduces the concept of *test stringency*, a test being most stringent if it assigns a minimal unconditional uncertainty safeguard in our sense of the term. Von Mises uses the expression *error chance* for what we call the uncertainty safeguard and *success rate* for what we denote by P_r , for which the expression *stochastic credibility* might be preferable in the common domain of test procedure and estimation.

REFERENCES

- Anscombe, F. J. (1951). *Mind*, **60**, 299.
 Columbia University Statistical Research Group (1945). "Sequential Analysis of Statistical Data: Applications". Report 255, Section 1, p. 13.
 Eisenhart, C., Hastay, M. W., and Wallis, W. A. (Eds.) "Selected Techniques of Statistical Analysis". McGraw-Hill, New York.
 Fisher, R. A. (1925). "Statistical Methods for Research Workers". Oliver and Boyd, Edinburgh. (11th ed., 1950).
 Greenwood, M. (1935). "Epidemics and Crowd Diseases". Williams and Norgate, London.
 Hill, A. Bradford (1951). *Brit. med. Bull.*, **7**, (No. 4), 278.
 Jackson, R. W. B. (1936). *Statist. Res. Mem.*, **1**, 138.
 Kendall, M. G. (1946). "Advanced Theory of Statistics", vol. 2. 2nd ed., pp. 270, 272. Griffin, London.
 Mackay, D. (1951). "Hospital Morbidity Statistics". *Studies on Medical and Population Subjects*, No. 4. H.M.S.O., London.
 Meredith, G. P. (1951). *Brit. J. Psychol.*, (Gen. Sect.), **42**, 322.
 von Mises, R. (1943). *Ann. math. Statist.*, **14**, 238.
 Neyman, J. (1937). *Philos. Trans. A.*, **236**, 333.
 — (1950). "First Course in Probability and Statistics". Holt, New York.
 — and Pearson, E. S. (1933). *Proc. Camb. Phil. Soc.*, **29**, 492.
 Wald, A. (1942). "On the Principles of Statistical Inference". *Notre Dame Mathematical Lectures*, No. 1. Indiana.
 — (1947). "Sequential Analysis", pp. 16–17. Wiley, New York.