ON THE PROBLEM OF TESTING HYPOTHESES

By R. v. Mises

Harvard University

1. Introduction. The following is known as the problem of testing a simple statistical hypothesis. The probability distribution of a variate X depends on a parameter ϑ. In the course of experiments each time a value x of X is observed, one pronounces one of the two assertions: "ϑ equals ϑ_0" or "ϑ is different from ϑ_0." The first assertion is made when the observed value x falls in a "region of acceptance" A, the second, if x falls in the complementary region $\bar{A}$. What is the chance of these assertions being correct and how can A be chosen to make this chance as high as possible?

The distribution for the variate X is considered as given. Let $P(x \mid \vartheta)$ be the probability of the value of X being $\leqq x$. It is obvious that to know $P(x \mid \vartheta)$ is not sufficient for computing the success or error chances of the above assertions. There is another distribution function $P_0(\vartheta)$ involved which we may call the initial or the a priori or the over-all distribution of the parameter ϑ. The meaning of $P_0(\vartheta)$ is as follows. In the infinite sequence of trials there will be among the first N experiences N_1 cases where the assertion that the parameter value is $\leqq \vartheta$ proves correct. Then $P_0(\vartheta)$ is the limit of the ratio N_1/N when N tends to infinity. If N_0 is the number of cases in which the actually pronounced assertions $\vartheta = \vartheta_0$ or $\vartheta \neq \vartheta_0$ respectively, prove correct, the limit of N_0/N is the success chance and of $1 - N_0/N$ the error chance of the test under consideration. It would not make any sense to assume that an error chance exists but the over-all chance $P_0(\vartheta)$ does not.[1]

The success and error chances for the assertions $\vartheta = \vartheta_0$ and $\vartheta \neq \vartheta_0$ depend on both functions $P(x \mid \vartheta)$ and $P_0(\vartheta)$. But in most practical cases nothing or very little is known about the parameter distribution. Usually, only the limits within which ϑ varies are known, or a set of distinct values is given which ϑ can assume. Therefore, the problem of testing a hypothesis must be modified in the following way. We ask: *What can be said about the error and success chances of the two alternative assertions and about the choice of the region of acceptance, if $P_0(\vartheta)$ is entirely or partly unknown?* This form of the question corresponds more or less to the conception generally adopted today.

In section 4 of this paper a complete answer to the question is presented for the case of a parameter distribution that is entirely unknown except for the range of possible ϑ-values. This solution, with the restriction to a parameter assuming distinct values only, was already given by Robert W. B. Jackson in a paper devoted mainly to some genetical problems [1]. The particular circumstances prevailing under the restriction to distinct parameter values will be discussed

[1] The expression "chance" rather than "probability" is used here since no randomness is required. Cf. the author's paper [2] p. 157.

in section 8. In section 6 the result is extended to composite hypotheses and in section 7 to problems in several dimensions. An important case of restrictions imposed to $P_0(\vartheta)$ is discussed in section 9.

In the preceding lines the subject of testing a statistical hypothesis was presented in its simplest form, with one scalar variate and one parameter, in order to discard all non-essential complications which would serve only to veil the principal point. For the same reason it is to be understood, in the following text, that region (in one dimension) will mean an interval or a finite number of intervals, and distribution will mean a set of concentrated values at distinct points with a continuous density in between or a continuous density throughout. If, for the sake of brevity, a Stieltjes integral is used, nothing else is meant than the combination of a sum and an ordinary integral of a continuous function. With respect to the parameter ϑ the distributions $P(x \mid \vartheta)$ are considered as either defined for distinct ϑ-values only or as continuous functions, etc.

2. Error chance. Success rate. J. Neyman who must be credited with successfully promoting many problems of mathematical statistics introduced the distinction between errors of first and second type and made this the basis of his approach in dealing with the theory of tests. An error of first kind is committed if the assertion $\vartheta \neq \vartheta_0$ is made when ϑ equals ϑ_0 ; an error of second kind occurs when the assertion $\vartheta = \vartheta_0$ proves incorrect.[2] The chances P_I and P_{II} of these two events can easily be computed, if the distributions $P(x \mid \vartheta)$ and $P_0(\vartheta)$ are considered as known. From $P(x \mid \vartheta)$ we derive the probability $P(A \mid \vartheta)$ for x falling in the region A. In particular $P(A \mid \vartheta_0)$ will be designated by $1 - \alpha$. Thus α is the probability of x falling in $\bar{A}$ when $\vartheta = \vartheta_0$. The function $P_0(\vartheta)$ can have, at the point $\vartheta = \vartheta_0$, a jump of magnitude π_0 . The set of all ϑ-values except ϑ_0 will be called $\bar{H}$. Then the two error chances are obviously

$$P_I = \alpha\pi_0 \qquad P_{II} = \int_{(\bar{H})} P(A \mid \vartheta)\, dP_0(\vartheta). \tag{1}$$

By the integral over $\bar{H}$ is meant that the term $P(A \mid \vartheta_0)\pi_0$ in the summation has to be omitted. The formulae (1) show anew that it would be senseless to speak of error chances without assuming that an over-all distribution $P_0(\vartheta)$ exists.

In all papers that follow Neyman's line of thought first and second type error chances are discussed. But the formulae (1) are seldom written down.[3] It is incorrect to say that α is the chance of a first type error and it is likewise incorrect to say that the chance of a second type error depends on ϑ; it depends on the distribution of ϑ.

The total error chance is

$$P_E = P_I + P_{II} = \alpha\pi_0 + \int_{(\bar{H})} P(A \mid \vartheta)\, dP_0(\vartheta) \tag{2}$$

[2] See e.g. ref. [4], [5] or various other publications by the same author.

[3] They are included e.g. in equation (1) of A. Wald's paper [5].

and $1 - P_E$ is the success chance. If the distribution $P(x \mid \vartheta)$, the region of acceptance A, and the test value ϑ_0 are given, P_E depends on $P_0(\vartheta)$ only. If we make $P_0(\vartheta)$ coincide successively with all functions not excluded by some preliminary knowledge about the over-all distribution, there must exist a definite least upper bound (l.u.b.) of P_E since P_E has the upper bound 1. The value

$$S = 1 - \text{l.u.b. } P_E$$

is the greatest lower bound of the success chance. In other words, for any positive ϵ there exists a $P_0(\vartheta)$ for which the success chance is $S + \epsilon$ and S is the greatest number for which this holds true. We therefore call S the *sure success rate* or, briefly, the *success rate* for the test under consideration. If the success rate S' for a region of acceptance A' is greater than S, the test using A' will be briefly called preferable to that using A.

Neyman's approach consists in comparing two regions A and A' with the same α. The difference of the respective error chances P_E and P'_E is according to (2):

$$P_E - P'_E = \int_{(\bar{H})} [P(A \mid \vartheta) - P(A' \mid \vartheta)]\, dP_0(\vartheta) \tag{3}$$

This difference is non-negative, whatever is taken for $P_0(\vartheta)$, if for all values of ϑ

$$P(A \mid \vartheta) \geqq P(A' \mid \vartheta). \tag{4}$$

In this case $P_E \geqq P'_E$ and l.u.b. $P_E \geqq$ l.u.b. P'_E and therefore $S \leqq S'$. If a region A' can be found for which (4) holds for whatever A, Neyman calls the test using A' a *most powerful test.* In fact, this test has at least as large a success rate as any other test using a region of acceptance with the same α. Neyman does not use the concept of success rate as introduced here, but implicitly the success chance is the criterion underlying his analysis of tests.[4]

The theory of most powerful tests would supply a complete solution of our problem, if (1) a most powerful test existed in all cases, i.e. for all distributions $P(x \mid \vartheta)$ and all ϑ_0 ; and if (2) a sufficient indication how to chose α were given. Unfortunately it turns out that in almost no practical case a region A' of this kind can be found. The various substitutes for a most powerful test as proposed by Neyman and others (unbiased test, test of type A, etc.) need not be discussed here, since it is obvious that nothing can be said about the difference $S - S'$, if (4) is not fullfilled for all A and ϑ. As to the choice of α, the expression

[4] This can be seen e.g. from the justification of most powerful tests as given by A. Wald [7] p. 15–16. Moreover, the recommendation of a test with highest success rate as the "best" (which is not the purpose of the present paper) could be justified from the standpoint of the general theory developed by Wald [6]. Wald introduces an arbitrary weight function for defining a "best" test. If the error weight is taken as one in the case of a false answer and as zero for each correct answer, Wald's "best" test coincides with the test of highest success rate. The present paper includes only statements that refer to the actual numbers of correct and false answers, independently of any arbitrary assumption about an error weight.

"level of significance" used by Neyman, leaves it open whether a high or a low value of α is preferable.

3. Preliminary example. Before attacking the general problem the discussion of a very simple example may provide some information. Let the distribution of the variate X be given by the density

$$p(x \mid \vartheta) = 1 + \vartheta^2(x^2 - \tfrac{1}{3}), \qquad 0 \leqq x \leqq 1. \tag{5}$$

It is immediately seen that the integral of p over the interval 0 to 1 equals 1 for each ϑ and that $p \geqq 0$, if ϑ lies in the limits $-\sqrt{3}$, $\sqrt{3}$. Let this be the only information we possess about the over-all distribution $P_0(\vartheta)$. The value to be tested may be $\vartheta_0 = 0$. The density for this parameter value reduces to $p(x \mid 0) = 1$ and thus the probability of x falling within the interval x_1, x_2 equals $x_2 - x_1$, if $\vartheta = \vartheta_0$. According to the notation introduced above we may consider as intervals of acceptance A all intervals with the limits x_1, $x_1 + 1 - \alpha$, where $0 \leqq x_1 \leqq \alpha$.

The function $P(A \mid \vartheta)$ is now given by

$$\begin{aligned} P(A \mid \vartheta) &= \int_{x_1}^{x_1+1-\alpha} p(x \mid \vartheta)\, dx \\ &= 1 - \alpha + (1 - \alpha)\vartheta^2\left[x_1^2 + x_1(1 - \alpha) - \frac{\alpha(2 - \alpha)}{3}\right]. \end{aligned} \tag{6}$$

In particular, for the interval A' between 0 and $1 - \alpha$:

$$P(A' \mid \vartheta) = 1 - \alpha - (1 - \alpha)\vartheta^2 \frac{\alpha(2 - \alpha)}{3}. \tag{7}$$

The difference of these two expressions is non-negative:

$$P(A \mid \vartheta) - P(A' \mid \vartheta) = (1 - \alpha)\vartheta^2 x_1(x_1 + 1 - \alpha) \tag{8}$$

Thus the interval 0, $1 - \alpha$ is seen to be a most powerful one. The error chance of this test is according to (2):

$$\begin{aligned} P_E' &= \alpha\pi_0 + \int_{(\bar{H})}\left[1 - \alpha - \vartheta^2(1 - \alpha)\frac{\alpha(2 - \alpha)}{3}\right]dP_0(\vartheta) \\ &= \alpha\pi_0 + (1 - \alpha)(1 - \pi_0) - (1 - \alpha)\frac{\alpha(2 - \alpha)}{3}\int_{(\bar{H})} \vartheta^2\, dP_0(\vartheta). \end{aligned} \tag{9}$$

The last integral is non-negative and can approach zero indefinitely since the total amount $1 - \pi_0$ can be concentrated at a point $\vartheta \neq 0$ with $\vartheta^2 < \epsilon$. Therefore the l.u.b. of P_E' for given α and π_0 is

$$\alpha\pi_0 + (1 - \alpha)(1 - \pi_0)$$

On the other hand, this is a linear function of π_0 which takes its extreme values at the ends of its interval, $\pi_0 = 0$ and $\pi_0 = 1$. Thus the larger of the two values

α and $1 - \alpha$ is the l.u.b. of P'_E, if $P_0(\vartheta)$ is subjected to no further restriction. The success rate of the test under consideration is accordingly the smaller of the two quantities α and $1 - \alpha$.

For $\alpha = 0.99$ or $\alpha = 0.01$ the success rate is 0.01. This means: If we use the most powerful test at a level of significance of either 99% or 1%, we risk in both cases that 99% of all assertions will be false. If $\alpha = \frac{1}{2}$, the success rate reaches its maximum value which is $\frac{1}{2}$ too. On the other hand it can be seen that each interval of length $\frac{1}{2}$ with not too large x_1 would lead to the same success rate. In fact, the error chance P_E for the interval x_1, $x_1 + 1 - \alpha$ is according to (9) and (6)

$$P_E = \alpha\pi_0 + (1 - \alpha)(1 - \pi_0) - (1 - \alpha)\left[\frac{\alpha(2 - \alpha)}{3} - x_1(x_1 + 1 - \alpha)\right]\int_{(\bar{H})} \vartheta^2\, dP_0(\vartheta). \tag{9'}$$

Therefore, the same reasoning as before applies, if the factor in brackets is non-negative. This is the case for $\alpha = \frac{1}{2}$ if the interval begins at a point $x_1 \leqq \frac{1}{4}(\sqrt{5} - 1) = 0.309$. Among these intervals, that with $x_1 = 0$ can be considered as preferable since its success chance for any $P_0(\vartheta)$ is at least as high as that of any other interval.

Now, let us assume that in the definition (5) of $P(x \mid \vartheta)$ the factor ϑ^2 is replaced by some function $g(\vartheta)$ which takes positive and negative values (within $-3/2$ and 3) while ϑ varies from $-\sqrt{3}$ to $\sqrt{3}$. Then equation (6) shows that for any two intervals of acceptance A and A' the difference $P(A \mid \vartheta) - P(A' \mid \vartheta)$ changes its sign at least once with varying ϑ. Thus no most powerful test interval exists. But, applying (9) and calling g_1 the (negative) minimum value of $g(\vartheta)$ we find now

$$\alpha\pi_0 + (1 - \alpha)(1 - \pi_0) - g_1(1 - \alpha)\left[\frac{\alpha(2 - \alpha)}{3} - x_1\,(x_1 + 1 - \alpha)\right](1 - \pi_0)$$

as the l.u.b. of the error chance of A' for given α and π. Thus the smaller of the quantities

$$1 - \alpha \quad \text{and} \quad 1 - (1 - \alpha)\left[1 - g_1\frac{\alpha(2 - \alpha)}{3}\right]$$

is the success rate of the test using A'. If g_1 is given we can find, by differentiation the value supplying the highest success rate. Using (9') instead of (9) we find in a similar way the success rates for any other interval. It turns out that $S = \frac{1}{2}$ for the interval extending from the above given value $x_1 = 0.309$ to 0.809.

There are three things we may learn from this example. (1) It can happen that a most powerful test, at a high or at a low level of significance, has an extremely poor success rate; (2) In the case where a most powerful test with the highest possible success rate exists, there may be other intervals with the same success rate; (3) If no most powerful test exists, there is no need to look

for some substitute definition; the success rate for any kind of test can be found independently of its being most powerful or not.

4. General solution for a simple hypothesis. The distribution $P(x \mid \vartheta)$ of the variate X, the parameter value ϑ_0 to be tested, and the set of all possible values of ϑ are supposed to be given. The set of all possible ϑ-values except ϑ_0 is called $\bar{H}$. Choose a region of acceptance A and compute first, for all ϑ, the magnitude

$$P(A \mid \vartheta) = \int_{(A)} dP(x \mid \vartheta). \tag{10}$$

In particular, the value of this integral for $\vartheta = \vartheta_0$ will be called $1 - \alpha$ and its maximum value or its l.u.b. on $\bar{H}$ will be denoted by β:

$$P(A \mid \vartheta_0) = 1 - \alpha, \qquad \text{l.u.b.}_{(\bar{H})}\, P(A \mid \vartheta) = \beta. \tag{11}$$

The chance of committing an error in asserting $\vartheta = \vartheta_0$ when x falls in A or $\vartheta \neq \vartheta_0$ in the case x falls in the complement $\bar{A}$ is according to (2)

$$P_E = \alpha\pi_0 + \int_{(\bar{H})} P(A \mid \vartheta)\, dP_0(\vartheta),$$

where π_0 is the jump of $P_0(\vartheta)$ at the abscissa $\vartheta = \vartheta_0$, or the a priori chance of ϑ_0. The domain of integration over $\bar{H}$ is $(1 - \pi_0)$ and therefore $\beta(1 - \pi_0)$ the l.u.b. of the integral. Thus[5]

$$\text{l.u.b. } P_E = \max \{\alpha\pi_0 + \beta(1 - \pi_0)\}.$$

As π_0 can take all values between zero and one, the lowest upper bound of P_E is either α or β. The success rate S, i.e. the greatest lower bound of $1 - P_E$, is consequently the smaller of the quantities $1 - \alpha$ and $1 - \beta$.

If the distribution $P(x \mid \vartheta)$ is given and a region of acceptance A for a test value ϑ_0 chosen, the success rate of this test equals the smaller of the two quantities

$$1 - \alpha = P(A \mid \vartheta_0) \quad \textit{and} \quad 1 - \beta = 1 - \text{l.u.b.}_{(\bar{H})}\, P(A \mid \vartheta), \tag{12}$$

if nothing is known about the initial distribution of the parameter except its range. Finding a region of acceptance, A, with the highest success rate, is then a simple maximum-minimum problem.

This solution is not restricted to some rarely occurring type of distributions $P(x \mid \vartheta)$ and it is insofar a complete one as it does not leave undetermined the value of α. Using Neyman's terminology we would have to say: The success rate is the smaller of the two quantities: 1 minus level of significance and minimum power of the test.

It follows from the definitions (12) that, if $P(A \mid \vartheta)$ is continuous in a ϑ-

[5] This formula was given by Jackson [1] p. 148 for the "case when the set of alternatives is discontinuous". Jackson calls the test with highest success rate a "most stringent test"

interval including ϑ_0, and ϑ is allowed to take all values of this interval, β cannot be smaller than $1 - \alpha$:

$$\beta \geqq 1 - \alpha \quad \text{or} \quad \alpha + \beta \geqq 1.$$

Thus $1 - \alpha$ and $1 - \beta$ cannot possibly both be greater than $\frac{1}{2}$. The greatest possible success rate is then $\frac{1}{2}$ and it can be reached only if $\alpha = \beta = \frac{1}{2}$. We state: *No test can have a success rate S greater than $\frac{1}{2}$, if ϑ can vary in an interval including ϑ_0 without any restriction and $P(A \mid \vartheta)$ is a continuous function of ϑ in this interval.*

We will see later, in sections 8 and 9, how certain restrictions imposed to $P_0(\vartheta)$ which are effective in some problems improve the success rate of a test.

5. Examples. Let us assume that the variate X is normally distributed according to

$$P(x \mid \vartheta) = \Phi[h(x - \vartheta)], \qquad \Phi(u) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{u} e^{-x^2}\, dx. \tag{13}$$

The parameter value to be tested may be taken as $\vartheta_0 = 0$ without loss of generality, since in all other cases $X - \vartheta_0$ can be considered as the variate. If the interval x_1, x_2 is chosen for the region of acceptance, we have

$$P(A \mid \vartheta) = \phi[h(x_2 - \vartheta)] - \phi[h(x_1 - \vartheta)]. \tag{14}$$

The right hand side becomes a maximum, if

$$\phi'[h(x_2 - \vartheta)] = \phi'[h(x_1 - \vartheta)], \qquad \text{i.e.} \quad \vartheta = \tfrac{1}{2}(x_1 + x_2).$$

Therefore, for $\vartheta_0 = 0$

$$1 - \alpha = \phi(hx_2) - \phi(hx_1), \qquad \beta = \phi(\tfrac{1}{2}h(x_2 - x_1)) - \phi(\tfrac{1}{2}h(x_1 - x_2)).$$

Both quantities have the value $\frac{1}{2}$, if and only if

$$x_1 = -x_2, \qquad \phi(hx_1) = \tfrac{1}{4}, \qquad \phi(hx_2) = \tfrac{3}{4}. \tag{15}$$

These are the *probable limits* of x. The conclusion is that *the probable limits supply the interval with the highest possible success rate $S = \frac{1}{2}$.*

The result is not restricted to the particular form of the function ϕ, it remains valid, if ϕ is replaced by any function whose derivative ϕ' has one maximum and decreases both ways symmetrically. It is well known that this test which has always been used by statisticians and is here proved to have the maximum success rate, is neither most powerful nor even, for a general ϕ, unbiased. We also see that the interval determined by (15) is the only closed interval with maximum success rate.

Our method supplies the analogous solution for the case of an unsymmetric distribution also. Assume the density

$$p(x \mid \vartheta) = f(x - \vartheta), \tag{16}$$

where $f(u)$ is supposed to have only one maximum, say at the point $u = 0$. The value to be tested may again be chosen as $\vartheta_0 = 0$. For the interval x_1, x_2 as region of acceptance we have

$$P(A \mid \vartheta) = \int_{x_1}^{x_2} f(x - \vartheta)\, dx = \int_{x_1-\vartheta}^{x_2-\vartheta} f(u)\, du.$$

The last expression becomes a maximum with respect to ϑ, if

$$f(x_1 - \vartheta) = f(x_2 - \vartheta).$$

The maximum will occur at the point $\vartheta = 0$ and accordingly coincide with $1 - \alpha$, if $f(x_1) = f(x_2)$. Thus we have a region of acceptance with the highest possible success rate $\frac{1}{2}$, if x_1, x_2 are determined by

$$\int_{x_1}^{x_2} f(u)\, du = \tfrac{1}{2}, \qquad f(x_1) = f(x_2). \tag{17}$$

Under the assumptions made for $f(u)$ there exists exactly one pair of values x_1, x_2 obeying these equations. This kind of test too has been much used by statisticians, but an account of its merits has so far not been given.

Another example is supplied by the density function

$$p(x \mid \vartheta) = \vartheta^2 x e^{-\vartheta x}, \qquad x \geqq 0, \quad \vartheta > 0. \tag{18}$$

We derive for an interval x_1, x_2

$$P(A \mid \vartheta) = \int_{x_1}^{x_2} p(x \mid \vartheta)\, dx = (\vartheta x_1 + 1)e^{-\vartheta x_1} - (\vartheta x_2 + 1)e^{-\vartheta x_2}.$$

If ϑ_0 is the value to be tested, we have

$$1 - \alpha = (\vartheta_0 x_1 + 1)e^{-\vartheta_0 x_1} - (\vartheta_0 x_2 + 1)e^{-\vartheta_0 x_2}. \tag{19}$$

One may ask for an interval x_1, x_2 with the success rate $S = \frac{1}{2}$. Then equation (19) must be fulfilled with $\alpha = \frac{1}{2}$ and, moreover, $P(A \mid \vartheta)$ must take its maximum value at $\vartheta = \vartheta_0$. This provides the second condition

$$\frac{\partial P(A \mid \vartheta)}{\partial \vartheta} = 0 \quad \text{at} \quad \vartheta = \vartheta_0, \quad \text{i.e.} \quad x_2^2 e^{-\vartheta_0 x_2} = x_1^2 e^{-\vartheta_0 x_1}. \tag{19'}$$

There exists, for each $\vartheta_0 > 0$, one and only one pair of values x_1, x_2 obeying the two equations (19) and (19').

In all these examples it turned out that at least one interval with the success rate $S = \frac{1}{2}$ (the highest value for a distribution continuous with respect to ϑ) exists. It seems that this is a common property of most usual distribution functions $P(x \mid \vartheta)$. But we can easily give an example where the greatest S, at least for a single interval as region of acceptance, is smaller than $\frac{1}{2}$. Assume

$$P(x \mid \vartheta) = x + \vartheta x(1 - x)(2\vartheta^2 x - 1), \qquad 0 \leqq x \leqq 1, \qquad -1 \leqq \vartheta \leqq 1, \tag{20}$$

and let $\vartheta_0 = 0$ be the value subjected to testing. For any interval beginning at x and extending to $x + 1 - \alpha$ we find

$$P(A \mid \vartheta) = 1 - \alpha + a\vartheta + b\vartheta^3 \quad \text{with} \quad a = (1-\alpha)(2x-\alpha), \tag{21}$$
$$b = 2(1-\alpha)(-3x^2 + 3\alpha x - \alpha^2 + \alpha - x).$$

It is a necessary condition for a test with $S = \frac{1}{2}$—in the case of a differentiable $P(A \mid \vartheta)$—that the derivative of $P(A \mid \vartheta)$ vanishes at $\vartheta = \vartheta_0$. Thus we must have

$$\frac{\partial P(A \mid \vartheta)}{\partial \vartheta} = a + 3b\vartheta^3 = 0 \quad \text{for} \quad \vartheta = 0. \tag{22}$$

This shows that $2x - \alpha$ must be zero or $x = \frac{1}{4}$. On the other hand, for $\alpha = \frac{1}{2}$, $x = \frac{1}{4}$ the formula for $P(A \mid \vartheta)$ becomes

$$P(A \mid \vartheta) = \tfrac{1}{2} + \tfrac{3}{16}\vartheta^3.$$

Thus P has an inflexion point at $\vartheta = 0$ and its maximum, β, must be greater than $\frac{1}{2}$. In the present example, as ϑ goes up to 1, we have $\beta = 11/16$ and the success rate is $S = 5/16$. This does not exclude that intervals with a success rate between $5/16$ and $\frac{1}{2}$ exist. E.g. for $x = 0.45$ and $\alpha = \frac{1}{2}$ one finds the maximum $\beta = 0.60$ and thus $S = 0.40$. The optimum interval can be found by differentiating the formula for $P(A \mid \vartheta)$ with respect to x and α.

Examples with the ϑ restricted to distinct values will be discussed in section 8.

6. Composite hypotheses. We have the problem of testing a composite hypothesis, if instead of one value ϑ_0 a region H of ϑ-values is given and the assertions to be made in the course of experiments are "ϑ belongs to H" or "ϑ does not belong to H." The solution developed in section 4 applies to this case almost without modification.

Again, let $P(A \mid \vartheta)$ be the probability of x falling in the region of acceptance A. By $\bar{A}$ and $\bar{H}$ we denote the regions complementary to A in the sample space and to H in the ϑ-space. Then the error chance is

$$P_E = \int_{(H)} [1 - P(A \mid \vartheta)]\, dP_0(\vartheta) + \int_{(\bar{H})} P(A \mid \vartheta)\, dP_0(\vartheta). \tag{23}$$

This is an obvious generalisation of (2). The equation expresses the fact that each time x falls in $\bar{A}$ and ϑ in H or x in A and ϑ in $\bar{H}$, an error is committed. Let us use the notations

$$\pi_0 = \int_{(\bar{H})} dP_0(\vartheta)$$
$$\alpha = \text{l.u.b. of } P(\bar{A} \mid \vartheta) \text{ for } \vartheta \text{ in } H \tag{24}$$
$$\beta = \text{l.u.b. of } P(A \mid \vartheta) \text{ for } \vartheta \text{ in } \bar{H}$$

Then the first of the two integrals in (22) cannot be greater than $\alpha\pi_0$ and the second not greater than $\beta(1 - \pi_0)$. On the other hand no lower upper bound exists for either of these integrals, if π_0 is given and $P_0(\vartheta)$ subjected to no other restriction.

As π_0 varies between 0 and 1, the expression

$$\alpha\pi_0 + \beta(1 - \pi_0)$$

has its extreme values at the points $\pi_0 = 0$ and $\pi_0 = 1$ and these values are α and β. Accordingly the greater of the quantities α and β is the l.u.b. of P_E and the success rate S equals the smaller of the two quantities $1 - \alpha$ and $1 - \beta$. If $P(A \mid \vartheta)$ is continuous with respect to ϑ, we have again $\beta \geqq 1 - \alpha$, thus α and β cannot be both smaller than $\frac{1}{2}$ and no S can become $> \frac{1}{2}$.

If the hypothesis that ϑ lies in H is tested by means of a region of acceptance A, the success rate of this test equals the smaller of the two quantities $1 - \alpha$ and $1 - \beta$ which are the minimum of $P(A \mid \vartheta)$ for ϑ-values in H and the minimum of $P(\bar{A} \mid \vartheta)$ for ϑ-values outside H. The task of finding the region A with highest success rate is thus reduced to a simple maximum-minimum problem.

As an example let us take the density function

$$p(x \mid \vartheta) = f(x - \vartheta), \tag{25}$$

where $f(u)$ has a maximum at $u = 0$ and drops on both sides symmetrically and monotonically towards zero. The hypothesis to be tested may be given as

$$-b \leqq \vartheta \leqq b.$$

We find, if the interval x_1, x_2 is taken for region of acceptance:

$$P(A \mid \vartheta) = \int_{x_1}^{x_2} f(x - \vartheta)\, dx = \int_{x_1-\vartheta}^{x_2-\vartheta} f(u)\, du. \tag{26}$$

This function of ϑ has its maximum at $\vartheta = \frac{1}{2}(x_1 + x_2)$ and drops symmetrically both sides. If $\frac{1}{2}(x_1 + x_2)$ is supposed to lie in the interval $(0, b)$ we find

$$1 - \alpha = \int_{x_1+b}^{x_2+b} f(u)\, du, \qquad \beta = \int_{x_1-b}^{x_2-b} f(u)\, du.$$

Both quantities reach the value $\frac{1}{2}$, if we choose $x_2 = -x_1 = a$ and take for a the uniquely determined solution of

$$\int_{-a+b}^{a+b} f(u)\, du = \int_{-a-b}^{a-b} f(u)\, du = \tfrac{1}{2}. \tag{27}$$

For this interval the success rate has its highest possible value $\frac{1}{2}$.

7. Case of n variates and k parameters. The analysis given in section 4 for a simple hypothesis and in 6 for a composite one extends immediately to the case where instead of one variate X and one parameter ϑ a group of n variates $X_1, X_2, \cdots, X_n$ and a group of k parameters $\vartheta_1, \vartheta_2, \cdots, \vartheta_k$ are in question.

The region of acceptance A is now a portion of the n-dimensional sample space, determined by an interval of a function $F(x_1, x_2, \cdots x_n)$. The hypothesis to be tested will consist in assuming that the point $\vartheta_1, \vartheta_2, \cdots \vartheta_k$ falls into a certain region H of the k-dimensional parameter space. The success rate of such a test is again the smaller of the numbers $1 - \alpha$ and $1 - \beta$ where α and β are defined in exactly the same way as in the preceding section. The minimum of $P(A \mid \vartheta)$ when the ϑ-values fall into H is called $1 - \alpha$, and the maximum of the same function for all ϑ-combinations belonging to the complementary region $\bar{H}$ is β.

If the test function $F(x_1, x_2, \cdots x_n)$ is known, the interval with the highest success rate, can be found on the same lines as in the case of one variate. In fact, the quantity F takes the place of x in the former analysis. If the interval thus found has the success rate $\frac{1}{2}$, we know that no other test exists which would have a higher success rate as long as nothing is known about the a priori distribution in the parameter space. If a certain $F(x_1, x_2, \cdots x_n)$ does not lead to an interval with success rate $\frac{1}{2}$, one may try another test function. In the most general case the test function F with the highest success rate would be found by solving the problem of calculus of variation that consists in maximizing $1 - \alpha$ and $1 - \beta$. As a rule such an elaborate analysis will not be necessary.

To ask that a test be a most powerful one is too much and too little. It is too much since such a test does not exist in most cases. It is too little because there can exist another test (on a different level of significance) with a considerably higher success rate. The correct description of a most powerful test is that such a test can be shown, in a simple way, to have no smaller success chance whatever $P_0(\vartheta)$ is than a group of other tests. If a most powerful test exists, it may be considered preferable to all other tests of the same success rate, but there is no reason why it should be considered more favorable than any test with higher success rate. As to unbiased tests, and other substitutes for most powerful tests, nothing at all can be said about their merits as compared with that of other tests.

A simple example for tests with the highest possible success rate in the case of several dimensions is the following. Assume a density function

$$p(x \mid \vartheta) = f(x_1 - \vartheta_1, x_2 - \vartheta_2, \cdots x_n - \vartheta_n) \tag{28}$$

where $f(u_1, u_2, \cdots u_n)$ depends on the absolute values $|u_1|, |u_2|, \cdots |u_n|$ only and decreases monotonically with increasing $u_1^2 + u_2^2 + \cdots u_n^2$ in all directions. The parameter point $\vartheta_1 = \vartheta_2 = \cdots \vartheta_n = 0$ is to be tested. Let $F(x_1, x_2 \cdots x_n)$ be a function likewise depending on $|x_1|, |x_2|, \cdots |x_n|$ only, vanishing at the origin, and monotonically increasing with $x_1^2 + x_2^2 + \cdots x_n^2$. Then the set of points for which

$$F(x_1, x_2, \cdots x_n) \leqq C \tag{29}$$

is a region of acceptance with success rate $\frac{1}{2}$, if C is chosen in such a way as to have

$$\int_{(F \leqq C)} f(x_1, x_2, \cdots x_n)\, dx_1\, dx_2 \cdots dx_n = \tfrac{1}{2}. \tag{30}$$

This applies e.g. to normal populations. The proof is obvious.

8. Distinct parameter values. Tests with higher success rate than $\frac{1}{2}$ can be found, if the parameter ϑ is restricted to a set of distinct values. Take for instance our first example in section 3 and assume that ϑ can only take the three values 0, ± 1. Then in the second expression (9) for the error chance the integral can not approach the value zero since the region $\bar{H}$ does not include the point $\vartheta = 0$. The minimum value of the integral is $(1 - \pi_0)$ and thus

$$P'_E \leqq \alpha\pi_0 + (1 - \alpha)\left[1 - \frac{\alpha(2 - \alpha)}{3}\right](1 - \pi_0). \tag{31}$$

The success rate is the smaller of the two quantities

$$1 - \alpha \quad \text{and} \quad 1 - (1 - \alpha)\left[1 - \frac{\alpha(2 - \alpha)}{3}\right] = 1 - \beta.$$

The best value of α is found by equating α and β. This gives about $\alpha = \beta = 0.436$ and the success rate $S = 0.564$, for the region of acceptance $x = 0$ to $x = 0.564$. Other intervals or sets of intervals can be examined in the same way.

A more impressive example is the following. We draw $n = 12$ times from an urn which contains three balls, black ones and white ones. The observed value x is the number of white balls drawn. The probability ϑ of getting a white ball in one experiment can have one of the four values 0, 1/3, 2/3, 1, and we want to test the hypothesis $\vartheta = \vartheta_0 = 1/3$. The probability distribution is given by

$$\pi(x \mid \vartheta) = C_n^x \vartheta^x (1 - \vartheta)^{n-x} \tag{32}$$

Let us choose the set of points $x = 1, 2, \cdots 6$ as region of acceptance. Then

$$P(A \mid \vartheta) = \sum_{x=1}^{6} C_n^x \vartheta^x (1 - \vartheta)^{n-x}. \tag{33}$$

This sum can be computed for the 4 possible ϑ-values:

$P(A \mid \vartheta) =$	0	0.926	0.178	0
for $\vartheta =$	0	1/3	2/3	1

Thus $1 - \alpha$ has the value 0.926 and β equals 0.178. The success rate is the smaller of the two quantities 0.926 and 0.822, thus $S = 0.822$. If we restrict the region of acceptance to the points $x = 1$ to 5, the values of $1 - \alpha$ and $1 - \beta$ become 0.815 and 0.934, thus the success rate $S = 0.815$. In the first case we have more than 82% chance of making a correct assertion, whatever the a priori probability of ϑ may be!

It is obvious that this result will become more and more strongly marked, if the number of observations increases. This is connected with the subject of the next section.

9. Asymptotically increasing success rate. It seems strange that in the case of a continuously varying parameter and a distribution $P(x \mid \vartheta)$ which is continuous with respect to ϑ no test can have a success rate $> \frac{1}{2}$. One has the feeling that something might happen in the continuous problems similar to what was the case in the example of section 8. On the other hand our proof that $S \leqq \frac{1}{2}$, in sections 4 and 6, is conclusive and it applies to problems in more than 1 dimension also. The answer is that in the kind of problems where a large number of observations is involved a definite restrictive assumption about the over-all distribution $P_0(\vartheta)$ is silently introduced.

The problems we have here in mind are connected with sequences of distributions of the form

$$P_n(x \mid \vartheta) = \phi_n(x - \vartheta), \tag{34}$$

where $\phi_1(u)$, $\phi_2(u)$, $\phi_3(u)$, $\cdots$ are cumulative distribution functions for distributions more and more concentrated around one point, say $u = 0$. In a rigorous form the sequence $\phi_n(u)$ can be described by the following statement: For each $\epsilon, \eta > 0$ exists a number $N(\epsilon, \eta)$ such that

$$\phi_n(\eta) - \phi_n(-\eta) \geqq 1 - \epsilon \qquad \text{for } n > N(\epsilon, \eta). \tag{35}$$

One wants to test the hypothesis

$$-b \leqq \vartheta \leqq b,$$

under the assumption that *the parameter distribution does not depend on n.* In this case, as we shall show, one can find for each $\epsilon > 0$ a region of acceptance A such that the success rate S_n of the test corresponding to this A and to $P_n(x \mid \vartheta)$ is greater than $1 - \epsilon$ for sufficiently large n.

We divide the region $\bar{H}$, i.e. $\mid \vartheta \mid > b$, into two parts $\bar{H}_1$ and $\bar{H}_2$ where $\bar{H}_1$ consists of the points $\mid \vartheta \mid \leqq b + 2\eta$ and satisfies the condition

$$\int_{(\bar{H}_1)} dP_0(\vartheta) \leqq \frac{\epsilon}{3}. \tag{36}$$

Then the region of acceptance will be

$$-a = -b - \eta \leqq x \leqq b + \eta = a,$$

and the probability of x falling in this region:

$$P_n(A \mid \vartheta) = \phi_n(b + \eta - \vartheta) - \phi_n(-b - \eta - \vartheta). \tag{37}$$

As long as ϑ belongs to H the right hand side in (37) is not smaller than $\phi_n(\eta) - \phi_n(-\eta)$ and thus, according to (35) the error chance of first kind

$$P_I^{(n)} = \int_{(H)} [1 - P_n(A \mid \vartheta)]\, dP_0(\vartheta) \leqq 1 - [\Phi_n(\eta) - \Phi_n(-\eta)] \leqq \frac{\epsilon}{2} \tag{38}$$

$$\text{for } n > N\left(\frac{\epsilon}{3}, \eta\right).$$

The error chance of second kind can be written as

$$P_{II}^{(n)} = \int_{(\bar{H}_1)} P_n(B \mid \vartheta)\, dP_0(\vartheta) + \int_{(\bar{H}_2)} P_n(A \mid \vartheta)\, dP_0(\vartheta). \tag{39}$$

The first of these integrals cannot be larger than $\frac{\epsilon}{3}$ according to (36) since $P_n(A \mid \vartheta) \leqq 1$. The second integral cannot exceed the maximum value of $P_n(A \mid \vartheta)$ for ϑ in $\bar{H}_2$. But if $\mid \vartheta \mid > b + 2\eta$ the two arguments of ϕ_n in (37) have always the same sign and are in absolute value greater than η. It then follows from (35), in connection with the fact that $\phi_n(u)$ increases monotonously from 0 to 1, that the difference of the two ϕ_n-values cannot exceed $\frac{\epsilon}{3}$ for $n > N(\epsilon/3, \eta)$. Therefore

$$P_{II}^{(n)} \leqq \frac{\epsilon}{3} + \frac{\epsilon}{3} \text{ and } S_n = 1 - P_I^{(n)} - P_{II}^{(n)} \geqq 1 - \epsilon \text{ for } n > N\left(\frac{\epsilon}{3}, \eta\right). \tag{40}$$

This result has a wide range of application in the cases where a hypothesis is tested on the basis of a large number of independent observations. Consider a sequence of variates $X_1, X_2, X_3, \cdots$ subject to probability distributions $Q_1(x_1), Q_2(x_2), Q_3(x_3), \cdots$. Let $x = F(x_1, x_2, \cdots x_n)$ be a *statistical function*, i.e. a function depending on the *distribution* of its n variables only, and ϑ the expected value of F. Then the general law of large numbers states that the distribution of x has the form (34) with ϕ_n satisfying the inequality (35), if the $Q_n(x)$ fulfill certain conditions concerning mainly their behaviour at infinity[6]. The proof of this theorem which is the real source of most "asymptotical" properties of statistical tests was given for the first time in 1936. The particular case where F is the arithmetical mean of the n variables $x_1, x_2, \cdots x_n$ has been known as Tchebychef's theorem since 1867.

Applying this general law of large numbers we can now state the following fact. *In testing a hypothesis about the expected value ϑ of any regular statistical function of n variates we can reach a success rate $1 - \epsilon$, no matter how small ϵ is, if the number n increases indefinitely and the initial distribution of ϑ is supposed to be independent of n.* On the other hand, no test with a success rate greater than $\frac{1}{2}$ is available, if an assumption of this type is not used.

[6] For exact conditions see ref. [3].

10. Summary. In this paper a solution of the problem of testing hypotheses is presented in the following sense. It is assumed that a probability distribution depending on some parameters is given and that nothing is known about the initial distribution of these parameters. For any simple or composite hypothesis about the parameters and any region of acceptance chosen in the sample space the success rate S is computed, i.e. the minimum chance for getting right answers out of the test. From the formulae given for S a test with highest success rate can easily be found in each case.

This theory shares the point of departure with the actually used theory which leads to the concept of most powerful tests. A most powerful test is described as a test which, by simple reasoning, can be seen to have no smaller success chance than any other test on the same "level of significance" α. In the rare cases where most powerful tests exist for all α-values, one of them, with an α-value singled out by our theory, has the highest success rate and then is preferable to all other tests which might have the same success rate. In all other cases our method supplies a test of highest success rate in no relation to "unbiased" tests or other current substitutes for most powerful tests.

Some of the main results are: No test has a success rate $>\frac{1}{2}$, if nothing is known about the parameters except the limits of their values and if the given distribution is a continuous function of the parameters. The success rate can be higher, if the parameters are restricted to certain distinct values. A success rate no matter how close to 1 can be reached in a sequence of tests based on an increasing number n of observations, if the initial distribution of the parameters is known to be independent of n.

REFERENCES

[1] ROBERT W. B. JACKSON, "Tests of statistical hypotheses in the case when the set of alternatives is discontinuous, illustrated on some genetical problems." *Stat. Res. Mem.*, Vol. 1 (1936), p. 138–161.

[2] R. v. MISES, "On the correct use of Bayes' formula," *Annals of Math. Stat.*, Vol. 13 (1942), p. 156–165.

[3] R. v. MISES, "Die Gesetze der grossen Zahl für statistische Funktionen," *Monatsh. Mathem. u. Physik*, Vol. 43 (1936), p. 105–128.

[4] J. NEYMAN, "Sur la vérification des hypothèses statistiques composées," *Bull. Soc. Math. de France*, Vol. 63 (1935), p. 246–266.

[5] J. NEYMAN, "Outline of a theory of statistical estimation based on the classical theory of probability," *Phil. Trans.*, Ser. A, Vol. 236 (1937), p. 333–380.

[6] A. WALD, "Contributions to the theory of statistical estimation and testing hypotheses," *Annals of Math. Stat.*, Vol. 10 (1939), p. 299–326.

[7] A. WALD, "On the principles of statistical inference," 1942, *Notre Dame Lect.* No. 1.