SOME PRINCIPLES OF THE THEORY OF TESTING HYPOTHESES[1]

By E. L. Lehmann

University of California, Berkeley

Introduction:

1. The likelihood ratio principle. The development of a theory of hypothesis testing (as contrasted with the consideration of particular cases), may be said to have begun with the 1928 paper of Neyman and Pearson [16]. For in this paper the fundamental fact is pointed out that in selecting a suitable test one must take into account not only the hypothesis but also the alternatives against which the hypothesis is to be tested, and on this basis the likelihood ratio principle is proposed as a generally applicable criterion. This principle has proved extremely successful; nearly all tests now in use for testing parametric hypotheses are likelihood ratio tests, (for an extension to the non-parametric case see [33]), and many of them have been shown to possess various optimum properties.

At least in the parametric case the likelihood ratio test has a number of desirable properties. Among these we mention:

(i) Frequently it is easy to apply and leads to a definite and reasonable test.
(ii) If the sample size is large, and if certain regularity conditions are satisfied an approximate solution can be given for the distribution problems that arise in the determination of size and power of the test (Wilks [32], Wald [25]). In fact, if the likelihood ratio is denoted by λ, $-2 \log \lambda$ approximately has a central χ^2-distribution under the hypothesis, a non-central χ^2-distribution under the alternatives. The number of degrees of freedom in these distributions equal the number of constraints imposed by the hypothesis.
(iii) As was shown by Wald [25], under certain restrictions the likelihood ratio test possesses various pleasant large sample properties.

In view of this, one may feel that the likelihood ratio principle, although perhaps not always leading to the optimum test, is completely satisfactory, and that a more systematic study of the problem of test selection is not necessary. Unfortunately, against the pleasant properties just mentioned there stands a very unpleasant one. Cases exist, in which the likelihood ratio test is not only unsatisfactory but worse than useless, and hence the likelihood ratio principle is not reliable. Examples of this kind were constructed independently by H. Rubin and C. Stein; the following is Stein's example.

---

[1] Parts of this paper were presented in an invited address at the meeting of the Institute of Mathematical Statistics on Dec. 30, 1948, in Cleveland, Ohio.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

Let X be a random variable capable of taking on the values 0, ± 1, ± 2 with probabilities as indicated:

	-2	2	-1	1	0
Hypothesis H:	$\frac{\alpha}{2}$	$\frac{\alpha}{2}$	$\frac{1}{2} - \alpha$	$\frac{1}{2} - \alpha$	α
Alternatives:	pC	$(1 - p)C$	$\frac{1 - C}{1 - \alpha}\left(\frac{1}{2} - \alpha\right)$	$\frac{1 - C}{1 - \alpha}\left(\frac{1}{2} - \alpha\right)$	$\alpha \frac{1 - C}{1 - \alpha}$

Here α, C are constants, $0 < \alpha \leqq \frac{1}{2}$, $\frac{\alpha}{2-\alpha} < C < \alpha$, and p ranges over the interval [0, 1].

It is desired to test the hypothesis H at significance level α. The likelihood ratio test rejects when $X = \pm 2$, and hence its power is C against each alternative. Since $C < \alpha$, this test is literally worse than useless, for a test with power α can be obtained without observing X at all, simply by the use of a table of random numbers. It is worth noting that the test, which rejects H when $X = 0$, has power $\alpha \frac{1 - C}{1 - \alpha} > \alpha$, so that a reasonable test of the hypothesis in question does exist.

The existence of such examples gives added importance to the problem of developing a systematic theory of hypothesis testing. It is the purpose of the present paper to give a brief survey of the work done on some aspects of such a theory and to indicate certain extensions and modifications of the existing theory. Some examples and applications will be considered. These will be restricted to parametric problems. For applications to testing non-parametric hypotheses see [12].

The results of sections 5 and 8 were obtained jointly by Gilbert Hunt and Charles Stein in 1945. They have not been published and were communicated to me by Professor Stein. I should like to express to him my gratitude for acquainting me with this material and for giving me permission to include it in this paper. I should also like to acknowledge my indebtedness to Professor Henry Scheffé who read the manuscript and made many helpful suggestions.

2. Formulation of the problem. The problem of testing a statistical hypothesis was formulated by Neyman and Pearson [18] as follows.

A random variable X is known to be distributed over a space $\mathfrak{X}$ according to some member of a family of probability distributions $\{P_\theta^X\}$, $\theta \epsilon \Omega$. It will be assumed here that there is specified an additive class $\mathfrak{B}$ of sets in $\mathfrak{X}$, and that the probability distributions P_θ^X are probability measures defined over $\mathfrak{B}$. All sets or real valued functions mentioned in this paper will be assumed measurable $\mathfrak{B}$ unless otherwise stated. If $B \epsilon \mathfrak{B}$, we shall write for the measure assigned to B by P_θ^X interchangeably $P_\theta^X(X \epsilon B)$, $P_\theta^X(B)$, and if there is no possibility of confusion, $P_\theta(B)$. Throughout most of the paper it will be assumed that the probability measures P_θ^X are absolutely continuous with respect to a given sigma finite measure μ defined over $\mathfrak{B}$, so that there exist non-negative functions f_θ such that

$$P_\theta(B) = \int_B f_\theta(x)\, d\mu(x). \tag{2.1}$$

We shall then say that $f_\theta(x)$ is a generalized probability density *w.r.* to μ.

A statistical hypothesis H specifies a subset ω of Ω, and states that the distribution of X is some P_θ^X with $\theta \in \omega$. A test of H is any subset w of $\mathfrak{X}$, the convention being that H is rejected if the observed value x of X is in w, and that in the contrary case H is accepted. The selection of w is to be made as follows. A number α is given, $0 < \alpha < 1$, the level of significance, and w must be such that

$$P_\theta(w) = \alpha \text{ for all } \theta \in \omega. \tag{2.2}$$

Subject to this restriction it is desired to maximize $P_\theta(w)$ for θ in $\Omega - \omega$. The interpretation of these conditions is immediate. Since $P_\theta(w)$ is the probability of rejecting H computed under the assumption that P_θ^X is the distribution of X, equation (2.2) states that the probability of rejecting H is to be α (usually some small number such as .01 or .05) whenever H is true. Similarly the second condition expresses the fact that H is to be rejected with high probability when θ is in $\Omega - \omega$.

Naturally the second condition is not to be taken literally but rather as a loosely stated principle of choice. For in general there will exist a unique set w maximizing $P_{\theta_1}(w)$ for any given $\theta_1 \in \Omega - \omega$, but this w will change with θ_1. The condition has a clear meaning only in the case that the set $\Omega - \omega$ contains only a single point, and in a few special problems in which the same set w maximizes $P_\theta(w)$ for all $\theta \in \Omega - \omega$. In the general case there are available two main methods for making the condition precise. One may restrict consideration to some class of "nice" tests, so that within this class the maximization of $P_\theta(w)$ can be achieved uniformly for $\theta \in \Omega - \omega$. Alternatively, instead of asking that a local optimum property hold uniformly, one may look for a test whose power function possesses some optimum property in the large. Both of these approaches have an element of arbitrariness. In the first, the selection of a class of nice tests, in the second, the choice of an appropriate optimum property. Fortunately, in a number of important special cases, both methods, for various reasonable definitions, lead to the same test.

Before proceeding with this development, we shall modify the formulation of the problem slightly. First, as has been pointed out by many writers, it seems more natural to replace (2.2) by

$$P_\theta(w) \leqq \alpha \text{ for all } \theta \in \omega. \tag{2.3}$$

Secondly, we shall permit "randomized" tests (see [11, 29]), that is, instead of demanding that the statistician decide for each value of x whether to accept or to reject H, we shall allow the possibility that for certain x the decision be reached by means of some chance device such as a table of random numbers. By a test of H we shall therefore mean a function ϕ from $\mathfrak{X}$ to the interval $[0, 1]$, with the convention that when x is the observed value of X some chance experiment with two possible outcomes R, $\bar{R}$ will be performed such that $P(R) = \phi(x)$, and that H will be rejected when the outcome is R and will otherwise be accepted. The case of a non-randomized test w clearly is obtained as a special case by taking for ϕ the characteristic function of the set w.

For a test ϕ the probability of rejection is given by

$$\int_{\mathfrak{X}} \phi(x)\, d P_\theta^X(x) = E_\theta \phi(X) \tag{2.4}$$

where E_θ denotes expectation computed with respect to the probability distribution P_θ^X. We therefore obtain the following formulation of the problem: To determine a test function ϕ $(0 \leqq \phi(x) \leqq 1)$ which maximizes $E_\theta\, \phi(X)$, the *power* of ϕ against the alternative θ, for θ in $\Omega - \omega$ subject to the condition

$$E_\theta \phi(X) \leqq \alpha \text{ for all } \theta \in \omega. \tag{2.5}$$

In this connection it is convenient to use the term "*level of significance*" for the preassigned number α, and to define the *size* of the test ϕ as

$$\sup_{\theta \in \omega} E_\theta \phi(X). \tag{2.6}$$

Except in the trivial case that there exists a test of size $< \alpha$ whose power is 1 against all alternatives, the size of any optimum test (in fact, of any admissible test) equals the level of significance.

3. Testing against a simple alternative. A complete solution of the problem formulated in the last section is available only in the case that ω and $\Omega - \omega$ each contains only a single point, that is, in the case that both the hypothesis and the alternative are simple. The solution is then given by the fundamental lemma of Neyman and Pearson [18], which we may state in the following slightly more complete form.

THEOREM 3.1. *Let*

$$P_\theta(A) = \int_A f_\theta(x)\, d\mu(x). \tag{3.1}$$

(a) *For testing the hypothesis H: $\theta = \theta_0$ against the alternative $\theta = \theta_1$ at level of significance α, there exists a number k and a test ϕ of size α such that*

$$\begin{aligned} \phi(x) &= 1 \quad \textit{when} \quad f_{\theta_1}(x) > k f_{\theta_0}(x), \\ \phi(x) &= 0 \quad \textit{when} \quad f_{\theta_1}(x) < k f_{\theta_0}(x). \end{aligned} \tag{3.2}$$

(b) *If $f_{\theta_0}(x)$ and $f_{\theta_1}(x)$ are $\neq 0$ for all x in $\mathfrak{X}$, then a test ϕ is most powerful for testing H against $\theta = \theta_1$ if and only if it satisfies* (3.2) *except possibly on a set of $\mu-$ measure* 0[2]. (Note that the number k of (3.2) is essentially unique).

[2] Throughout the paper we shall consider two tests as equal if they differ only on a set of μ-measure 0.

The second half of the theorem may be paraphrased by saying that under the conditions stated the most powerful test is uniquely determined by (3.2) except on the set on which

$$f_{\theta_1}(x) = k\, f_{\theta_0}(x). \tag{3.3}$$

On this set the value of ϕ may be assigned arbitrarily provided the resulting test has size α. If in particular the set on which (3.3) holds has measure 0, the most powerful test is unique.

It should be mentioned that (3.1) is no restriction since any two probability measures P_1, P_2 defined over a common additive class can be represented in this form with $\mu = P_1 + P_2$. If the assumption of (b) is not satisfied, the theorem is still true in essence but some trivial modifications are necessary.

No such complete solution is available for the problem of testing a composite hypothesis against a simple alternative. However, as was shown in [11], this problem may in many cases be reduced to the one just considered. Let the hypothesis state that θ is an element of ω, and consider the simple alternative $\theta = \theta_1$. Suppose that an additive class of sets has been defined on ω (in most of the applications ω is a subset of Euclidean space, and the additive class is formed by the Borel sets contained in ω). Then for any probability distribution λ over ω,

$$h_\lambda(x) = \int_\omega f_\theta(x)\, d\lambda(\theta) \tag{3.4}$$

is a probability density function with respect to μ.

Under certain conditions to be stated below, the most powerful test ϕ_λ for testing the simple hypothesis H_λ that X is distributed with probability density h_λ against the alternative f_{θ_1} is also most powerful for testing the original hypothesis H against the same alternative. This is essentially the Bayes approach developed by Wald for his general decision theory, and in fact, under the conditions which we shall state, λ is a least favorable distribution over ω in the following sense. Let β_λ be the power of ϕ_λ against f_{θ_1}, and for any distribution λ^* over ω denote by H_{λ^*}, ϕ_{λ^*}, β_{λ^*} the associated hypothesis, the most powerful test for testing it against f_{θ_1}, and the power of this test respectively. Then λ is said to be least favorable if for all λ^*

$$\beta_\lambda \leqq \beta_{\lambda^*}. \tag{3.5}$$

THEOREM 3.2. *Suppose there exists a probability distribution λ over ω such that the most powerful test ϕ_λ of size α for testing H_λ against f_{θ_1} is of size α also with respect to the original hypothesis H. Then*

(i) *ϕ_λ is most powerful for testing H against f_{θ_1};*

(ii) *λ is a least favorable distribution.*

Also, if ϕ_λ is the unique most powerful test for testing H_λ against f_{θ_1}, it is the unique most powerful test for testing H against f_{θ_1}.

These results are essentially contained in Wald's work (see for example theorem 4.8 of [26]).

There are many trivial applications of this theorem to finding most powerful tests of one-sided hypotheses concerning a single real-valued parameter, such as testing $H: p \leqq p_0$ against $p = p_1(p_0 < p_1)$ when X has a binomial distribution with parameter p. As is well known, it turns out in a number of these cases that the most powerful tests are in fact uniformly most powerful against the one-sided class of alternatives.

In [11] Theorem 3.2 was used to determine most powerful tests of certain hypotheses concerning normal distributions. As an example consider the case that $X_1, \cdots, X_n$ are independently normally distributed with common mean ξ and variance σ^2. Denote by H_1 and H_2 the hypotheses $\sigma = 1$ and $\xi = 0$ respectively, and let the alternative be: $\xi = \xi_1$, $\sigma^2 = \sigma_1^2$. Then the most powerful test of H_1 rejects if

$$
\begin{aligned}
&\Sigma(x_i - \xi_1)^2 < k_1 \quad \text{when} \quad \sigma_1 < 1,\\
&\Sigma(x_i - \bar{x})^2 > c_1 \quad \text{when} \quad \sigma_1 > 1,
\end{aligned}
\tag{3.6}
$$

and accepts otherwise. Here k_1 and c_1 depend only on the level of significance, that is, are independent of ξ_1, σ_1. If $\xi_1 > 0$, the most powerful test for testing H_2 rejects if

$$
\begin{aligned}
&\Sigma(x_i - b)^2 \leqq k_2 b^2 \quad \text{when} \quad \alpha < \tfrac{1}{2},\\
&\frac{\bar{x}}{\sqrt{\Sigma(x_i - \bar{x})^2}} \leqq c_2 \quad \text{when} \quad \alpha \geqq \tfrac{1}{2},
\end{aligned}
\tag{3.7}
$$

and accepts H_2 otherwise. Here k_2 and c_2 depend only on α, while b depends on ξ_1, σ_1 and α.

These results indicate that even when the class of alternatives is larger than in the above problems, some improvement over the standard tests may be possible provided good power is desired only against a narrow class of alternatives.

4. Sufficient statistics. Before treating the problem of composite alternatives, we shall consider an important simplification that can be obtained by making use of sufficient statistics. This notion was introduced by R. A. Fisher, and was further developed by J. Neyman [13] and in [2] and [10]. Consider any measurable partition of $\mathfrak{X}$. For any point x in $\mathfrak{X}$, let $t(x)$ be that set of the partition in which x lies. A set in the range of t is said to be measurable if the corresponding set of points x is an element of $\mathfrak{B}$. Denote the class of measurable t-sets by $\mathfrak{A}$. Then the *statistic* $T = t(X)$ is a random variable defined over $\mathfrak{A}$. Kolmogoroff has shown how for any $B \in \mathfrak{B}$ one can define the conditional probability $P(B \mid t)$ of B given $T = t$ uniquely up to a set of measure zero by the equation

$$
P(B \cap t^{-1}(A)) = \int_A P(B \mid t)\, dP^T(t) \quad \text{for all} \quad A \in \mathfrak{A}. \tag{4.1}
$$

Suppose now that we are given a class $\mathfrak{F}$ of probability distributions for X, $\mathfrak{F} = \{P_\theta^X\}$, $\theta \in \Omega$. Denote by $P_\theta(B \mid t)$ the conditional probability of B given

$T = t$ computed for the distribution P_θ^X. The statistic T is said to be a *sufficient statistic* for $\mathfrak{F}$ (or for θ) if for every $B \in \mathfrak{B}$ there exists a determination of $P_\theta(B \mid t)$ that is independent of θ.

According to the above definition of statistic, $t(x)$ is an element of a measurable partition. However, one may consider instead any function t^* for which $t^*(x) = t^*(x')$ if and only if $t(x) = t(x')$, that is, any function that leads to this partition; the values that the function takes on are really immaterial. It will be convenient here to use this wider definition of statistic. For a rigorous treatment of some of the problems that will be referred to one needs to define an equivalence of statistics and to include in this definition the appropriate nullset considerations. A detailed account of these matters is given in [2] and [10].

From our present point of view tests are compared solely in terms of their power functions. On this basis two tests ϕ_1 and ϕ_2 may be considered equivalent if they have identical power, that is, if

$$E_\theta\phi_1(X) = E_\theta\phi_2(X) \text{ for all } \theta \in \Omega. \tag{4.2}$$

We can then state

THEOREM 4.1. *If T is a sufficient statistic for θ and $\phi(X)$ any test of a hypothesis concerning θ then there exists an equivalent test that is a function of T only.*

The proof of this theorem is immediate since

$$\psi(T) = E[\phi(X) \mid T] \tag{4.3}$$

is such a test.

It follows from Theorem 4.1 that we lose nothing by restricting consideration to tests based on a sufficient statistic.[3] The problem of determining whether or not some statistic is sufficient for a given family of distributions is simplified through the use of a criterion for sufficiency that can be checked on sight. This criterion is due to Neyman [13] who proved it in a somewhat special setting, and was recently proved in a very general form by Halmos and Savage [2]. It states that if $\mathfrak{F} = \{p_\theta\}$, $\theta \in \Omega$ is a family of generalized probability densities for X, then under certain mild restrictions a necessary and sufficient condition for $T = t(X)$ to be a sufficient statistic for $\mathfrak{F}$ is that $p_\theta(x)$ factors into one factor depending on θ but on x only through $t(x)$ and a second factor depending only on x.

The question arises as to which of various sufficient statistics to use. Since the purpose of introducing sufficient statistics is to reduce the complexity of a given statistical problem, one is led to seek a sufficient statistic that reduces the problem as far as possible and hence to the notion of a *minimal sufficient statistic*, a sufficient statistic T being *minimal* if it is a function of every other sufficient statistic (see [10]). It can be shown under fairly general conditions that a minimal sufficient statistic exists, and one can give an explicit construction for it.

---

[3] A justification for the use of sufficient statistics in the general statistical decision problem was given in [2].

As one would expect it turns out that the sufficient statistics commonly associated with various families of distributions are actually minimal. Thus for example, if $X_1, \cdots, X_n$ are independently normally distributed with common mean ξ and variance σ^2, the statistic $(\bar{X}, \Sigma(X_i - \bar{X})^2)$ is a minimal sufficient statistic for $\theta = (\xi, \sigma^2)$. If $X_1, \cdots, X_n$ are independently uniformly distributed over $(0, \theta)$, $\max(X_1, \cdots, X_n)$ is the minimal sufficient statistic for θ. If $\mathfrak{F}$ is the family of distributions according to which $X_1, \cdots, X_n$ are identically independently distributed according to an arbitrary univariate distribution (or according to an arbitrary probability density with respect to a fixed univariate measure), then the minimal sufficient statistic is obtained by defining for each point $x = (x_1, \cdots, x_n)$ the set $t(x)$ as the set of points obtainable from x by permutation of coordinates. Alternatively one can define it by $t(x_1, \cdots, x_n) = (\Sigma x_i, \Sigma x_i^2, \cdots, \Sigma x_i^n)$.

5. The principle of invariance. The notion of invariance was introduced into the statistical literature in the writings of R. A. Fisher, Hotelling, Pitman [20] and others, in connection with various special problems. A general formulation was given by Hunt and Stein who, in an unpublished paper [5], utilized this notion to find most stringent tests, and who obtained the examples of uniformly most powerful invariant tests that will be given below. The point of view in the present section is different from theirs however, since here invariance will only be considered as an intuitively appealing restriction that one may wish to impose on statistical tests.

We shall begin by considering an example. Suppose it were known that the height of people is distributed about a known mean, which for convenience we shall take to be zero, either according to a normal or to a Cauchy distribution, with unknown scale factor so that either

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}\,\theta} \exp\left(-\frac{x^2}{2\theta^2},\right) \qquad 0 < \theta < \infty \tag{5.1}$$

or

$$f_\theta(x) = \frac{\theta}{\pi} \frac{1}{\theta^2 + x^2}, \qquad 0 < \theta < \infty. \tag{5.2}$$

Suppose we wish to test from a sample $X_1, \cdots, X_n$ the hypothesis H that the true probability density belongs to the first of these classes against the alternative that it belongs to the second. Then it seems desirable that the decision of whether or not to accept H should be independent of the scale adopted for measuring the heights. For otherwise one worker expressing his data in feet might reject H while another worker using the same data but expressing them in inches would reach the contrary decision (In this connection see for example [34], p. 104). A "nice" test function ϕ therefore would be independent of the choice of scale, i.e., it would satisfy the condition

$$\phi(cx_1, \cdots, cx_n) = \phi(x_1, \cdots, x_n) \text{ for all } c > 0 \text{ and for all } (x_1, \cdots, x_n) \tag{5.3}$$

except possibly on a set N, independent of c and of measure zero.

On analyzing this problem one is led to the following observation. Multiplying each of the random variables $X_1, \cdots, X_n$ by the same constant leaves both ω and $\Omega - \omega$ invariant, i.e., if the X's are normally distributed with zero mean and arbitrary scale so are $cX_1, \cdots, cX_n$, and analogously for the Cauchy distributions. It is this fact that makes it so desirable to have ϕ invariant under multiplication of the x's by a common constant.

More generally consider measurable 1:1 transformations g of $\mathfrak{X}$ into itself, and let $Y = gX$. Suppose that when X is distributed according to $\theta \in \omega$, Y is distributed according to $\theta' \in \omega$—we shall then write $\theta' = \bar{g}\theta$—and that as θ ranges over ω so does θ'. Suppose that the analogous condition is satisfied for $\Omega - \omega$, so that the problem of testing ω against $\Omega - \omega$ is left invariant under g. Now whether one expresses the observations in terms of X or in terms of Y is essentially a matter of choice of coordinates. The principle of invariance asks that if such a change of coordinates leaves the problem invariant, then it should also leave the test invariant, i.e., if G is a group of measurable 1:1 transformations of $\mathfrak{X}$ such that

(5.4) $$\bar{g}\omega = \omega \text{ and } \bar{g}(\Omega - \omega) = \Omega - \omega \text{ for all } g \in G,$$

then ϕ should satisfy the condition

(5.5) $$\phi(gx) = \phi(x) \text{ for all } g \in G,$$

and for all x except on a set N independent of g and such that $\mu(N) = 0$. If this condition were not satisfied, two workers, using the same data but expressing them in different coordinate systems might arrive at contrary conclusions.

As an example consider the general linear univariate hypothesis. In canonical form $X_1, \cdots, X_r; X_{r+1}, \cdots, X_s; X_{s+1}, \cdots, X_n$ are independently normally distributed with common variance. The means of the first s variables are unknown, the means of the last n-s variables are known to be zero. The hypothesis states that the first r means are zero. Adding arbitrary constants to each of the variables of the middle group leaves ω and $\Omega - \omega$ invariant. So does any orthogonal transformation of the first r variables, and any orthogonal transformation of the last $n - s$ variables. Finally, the problem is also left invariant when all of the variables are multiplied by the same constant. It is easy to see that a function ϕ is invariant under these transformations if and only if it is a function of

$$\sum_{i=1}^{r} x_i^2 \Big/ \sum_{i=s+1}^{n} x_i^2 .$$

But, as is well known and easy to show, among all tests based on this statistic there is a uniformly most powerful one, namely the test that rejects H when

$$\sum_{i=1}^{r} x_i^2 \Big/ \sum_{i=s+1}^{n} x_i^2$$

is too large. Therefore, among all tests satisfying the condition of invariance the standard test is uniformly most powerful.

To formulate a corresponding reduction procedure in general, we define a function h on $\mathfrak{X}$ to be maximal invariant (under G) if it is invariant and if $h(x') = h(x)$ implies the existence of $g \in G$ such that $x' = gx$. Then a function φ on $\mathfrak{X}$ is invariant under G if and only if it depends on x only through $h(x)$, that is, if there exists a function ψ such that $\varphi(x) = \psi[h(x)]$. Hence a necessary and sufficient condition for a test to be invariant under G is that it be based on the statistic $Y = h(X)$. The principle of invariance therefore reduces the problem from X to $Y = h(X)$. To determine the resulting statistical reduction, that is, the simplification of the parameter space, one may consider the group $\bar{G}$ of transformations over Ω induced by G. If $v(\theta)$ is a maximal invariant function under $\bar{G}$, it is easily shown that the distribution of Y depends only on $v(\theta)$. Hence under the principle of invariance any two θ-values with common $v(\theta)$ (that is, such that each can be obtained from the other by a transformation of $\bar{G}$) are identified. If in particular $v(\theta)$ is constant over ω, the hypothesis H, when expressed for Y, becomes simple, and there may even exist a uniformly most powerful invariant test.

Besides for the example already mentioned this is the case for Hotelling's T^2-problem and for the hypothesis specifying the value of a multiple correlation coefficient. Another example is obtained when $X_1, \cdots, X_n$ are independently identically distributed, each with probability density $p_\theta(x)$ where under H_i $p_\theta(x) = f_i(x - \theta)$, $(i = 0, 1)$, and where it is desired to test H_0 against H_1. One may also in this example replace the location parameter by a scale parameter or have both parameters present.

It may be worth noting that the likelihood ratio test is invariant under any transformation leaving the statistical problem invariant. In the problems concerning normal distributions mentioned above, when there exists a uniformly most powerful invariant test, it coincides with the likelihood ratio test. That this is not so in general can be seen from Stein's example given in section 1. There the problem remains invariant under multiplication of X by -1, and there exists a uniformly most powerful invariant test. However, the likelihood ratio test is instead uniformly least powerful.

For certain applications it is more useful to consider a somewhat weaker definition of invariance. We shall say that a function φ is *almost invariant* under a group G of transformations if for each $g \in G$, $\varphi(gx) = \varphi(x)$ for all x except on a set N_g such that $\mu(N_g) = 0$. This definition differs from the previous one in that the null set N_g is now permitted to depend on g. It was shown by Hunt and Stein that under certain conditions on G, which are satisfied for the problems mentioned above, any almost invariant test is invariant.

We have indicated how for certain hypotheses one can find a group of transformations leaving the problem invariant, such that among all tests invariant under this group there exists a uniformly most powerful one. The question may be raised whether this approach is consistent, or whether there may exist some other group of transformations also leaving the problem invariant but leading to a different test. Also in problems where among all invariant tests there does not exist a uniformly most powerful one, the question arises whether one is using the totality of transformations leaving the problem invariant, or whether perhaps one can reduce the problem further. It therefore seems of interest to determine the totality of transformations leaving a given problem invariant. This was carried out for a few simple problems in [8].

We finally mention a connection between the notions of invariance and sufficiency. Consider any problem in which the variables $X_1, \cdots, X_n$ are independently identically distributed under all distributions of Ω. Such a problem clearly is left invariant under any permutation of the variables. Actually, these transformations leave not only ω and $\Omega - \omega$ invariant but each point of Ω individually. No essential reduction of the problem is obtained since the maximal invariant statistic is a sufficient statistic. It is easily seen that this will always be the case when the transformations leave Ω pointwise invariant, but that in this way one does not obtain all sufficient statistics. These can be obtained, however, by considering more general transformations, where each point x of $\mathfrak{X}$ is transformed into the points of $\mathfrak{X}$ according to a probability distribution P_x.

6. The principle of unbiasedness. As a second principle of reduction we shall consider the principle of unbiasedness proposed by Neyman and Pearson. A test is said to be unbiased [19] if

$$P_\theta \text{ (rejecting } H) \geq \alpha \text{ for all } \theta \in \Omega - \omega.$$

This seems a desirable property for a test to have since it assures that there do not exist θ_0 in ω and θ_1 in $\Omega - \omega$, for which

$$P_{\theta_0} \text{ (rejecting } H) > P_{\theta_1} \text{ (rejecting } H).$$

We shall therefore be concerned in this section with the totality of tests ϕ for which

$$\begin{aligned} E_\theta \phi(X) &\leq \alpha \quad \text{for all} \quad \theta \in \omega \\ E_\theta \phi(X) &\geq \alpha \quad \text{for all} \quad \theta \in \Omega - \omega. \end{aligned} \tag{6.1}$$

For a number of important special cases there exists, among all tests satisfying (6.1), one that is uniformly most powerful in $\Omega - \omega$ and uniformly least powerful in ω. (The latter property is of course very desirable since when H is true one wants to reject it as rarely as possible.) This follows immediately from well known results concerning best similar tests since for the problems in question Ω is a subset of a Euclidean space and for any test ϕ, $E_\theta\phi(X)$ is a continuous function of θ. If then Λ is the set of points that are boundary points both of ω and of $\Omega - \omega$, it follows from (6.1) that

$$E_\theta\phi(X) = \alpha \text{ for all } \theta \in \Lambda, \tag{6.2}$$

i.e., that ϕ is similar for θ in Λ. But if among all tests satisfying (6.2) there exists one that is uniformly most powerful in $\Omega - \omega$ and uniformly least powerful in ω, it automatically satisfies (6.1) as is seen by comparison with the test $\phi(X) \equiv \alpha$.

As an example suppose that $X_1, \cdots, X_n$ are independently normally distributed with common mean ξ and common variance σ^2. If the hypothesis is $H_1 : \sigma \leq 1$ and the alternatives are $\sigma > 1$, the set Λ becomes the line $\sigma = 1$. As was shown by Neyman and Pearson [18], among all tests satisfying (6.2) with this Λ, the test that rejects H_1 when $\Sigma(x_i - \bar{x})^2 \leq k$ (where k is an appropriately chosen constant) is uniformly most powerful for θ in $\Omega - \omega$, and uniformly least powerful for θ in ω.

If instead we consider testing the hypothesis $H_2 : \sigma = 1$ against the alternatives $\sigma \neq 1$, we find that $\Lambda = \omega$, and our problem reduces to that of finding the best test among all those that are similar in ω and unbiased. As is well known, it turns out that rejecting when $\Sigma(x_i - \bar{x})^2 \leq k_1$ and when $\Sigma(x_i - \bar{x})^2 \geq k_2$ (where $k_1 < k_2$ are two appropriately chosen constants) is uniformly most powerful among all similar unbiased tests.

A third hypothesis concerning σ that might be of interest is $H_3 : \sigma_1 \leq \sigma \leq \sigma_2$. Here Λ consists of the two lines $\sigma = \sigma_1$ and $\sigma = \sigma_2$ and it is easy to show that the test that is uniformly most powerful in $\Omega - \omega$ and uniformly least powerful in ω rejects H_3 if and only if $\Sigma(x_i - \bar{x})^2 \leq c_1$ or $\Sigma(x_i - \bar{x})^2 \geq c_2$ where again $c_1 < c_2$ are two appropriately selected constants.

The question arises as to the connection of the principles of invariance and unbiasedness. Clearly if there exists a unique test ϕ that is uniformly most powerful unbiased, this test is invariant under any group G leaving the problem invariant. If then in addition there exists a uniformly most powerful invariant (under G) test, this must coincide with ϕ. Thus, if both principles lead to a unique optimum solution, these solutions coincide.

We have seen that frequently optimum unbiased tests can be obtained through a study of tests that are similar over certain sets in the parameter space. The totality of similar tests was obtained for a number of important problems by Neyman and Pearson. In his 1937 paper on confidence intervals [15] Neyman gave a general method for constructing similar regions with the help of sufficient statistics. Let T be a sufficient statistic for $\theta \in \Lambda$. The condition for ϕ to be similar with respect to Λ and of size α, is that

$$E_\theta\phi(X) = E_\theta E[\phi(X) \mid T] = \alpha \text{ for all } \theta \in \Lambda, \tag{6.3}$$

i.e., that

$$E_\theta\{E[\phi(X) \mid T] - \alpha\} = 0 \text{ for all } \theta \in \Lambda. \tag{6.4}$$

Clearly any test ϕ for which

$$E[\phi(X) \mid t] = \alpha \text{ for almost all } t \tag{6.5}$$

is similar. This is the construction given by Neyman, and we shall say that a test ϕ satisfying (6.5) has the Neyman structure with respect to T. The question whether this exhausts the totality of similar tests is easily reduced to an analytic problem the solution of which is known in many special cases. This method was first employed by P. L. Hsu [3] for some problems concerning normal distributions, and was extended to other cases in [7]. The present general formulation was given by H. Scheffé and the author in [9] and [10]. We shall say that a family of distributions $\{P_\theta^T\}$, $\theta \in \Lambda$, is boundedly complete if

(i) $f(t)$ is bounded,

(ii) $E_\theta f(T) = 0$ for all $\theta \in \Lambda$

imply $f(t) = 0$ except on a set N with $P_\theta(N) = 0$ for all $\theta \in \Lambda$. Then we can state

THEOREM 6.1. *A necessary and sufficient condition for the totality of tests similar for Λ to have Neyman structure with respect to a sufficient statistic T is that $\{P_\theta^T\}$, $\theta \in \Lambda$, be boundedly complete.*

7. Tests whose power increases with the distance from the hypothesis. Frequently, even among the unbiased tests, there does not exist a uniformly most powerful one. The general univariate linear hypothesis with more than one constraint is an example of this situation. The following extension of the idea of unbiasedness may then be used to reduce the class of tests still further. Unbiasedness distinguishes between values of θ as they belong to ω or $\Omega - \omega$. However, one may further classify the points of $\Omega - \omega$ according to their "distance" from ω, and then ask of a test φ that the further be θ from ω the larger be the power $\beta_\varphi(\theta)$.

One possible such ordering of the alternatives is that induced by the envelope power function. Here the envelope power at θ (Wald [24]) is defined by

$$\beta_\alpha^*(\theta) = \sup_{\varphi \in \mathfrak{F}(\alpha)} \beta_\varphi(\theta) \tag{7.1}$$

where $\mathfrak{F}(\alpha)$ is the class of all tests φ with $E_\theta\varphi(X) \leq \alpha$ for all $\theta \in \omega$. Of two points θ_1, θ_2 one may then say that θ_1 is closer to ω than θ_2, equally close or less close, as $\beta_\alpha^*(\theta_1)$ is less than, equal to or greater than $\beta_\alpha^*(\theta_2)$. The distance of θ from ω is thus measured by the ease with which one can detect that the hypothesis is false when θ is the true parameter value.

When θ lies in a Euclidean space and $\beta_\varphi(\theta)$ is a continuous function of θ for all θ, as is the case in most applications, the condition that the power increase with β_α^* will usually imply that $\beta_\varphi(\theta_1) = \beta_\varphi(\theta_2)$ whenever $\beta_\alpha^*(\theta_1) = \beta_\alpha^*(\theta_2)$. In the case of the general linear hypothesis considered in section 5, for example, one would obtain the condition that the power be a function only of $\sum_{i=1}^{r} \xi_i^2/\sigma^2$ where $\xi_i = E(X_i)$. As was shown by P. L. Hsu [3], the standard (likelihood ratio) test is uniformly most powerful among all tests satisfying this condition. Analogous remarks apply to Hotelling's T^2- problem, and to the hypothesis specifying the value of the multiple correlation coefficient. The corresponding optimum properties in these cases were proved by Simaika [21].

It is interesting to compare the above condition with that of invariance.

This comparison yields nothing of interest if the totality of tests is considered. We may, however, restrict our attention to tests depending only on a sufficient statistic T. We already know that $\varphi(X)$ and $E[\varphi(X) \mid T]$ have identical power. In order to validate the comparison we wish to make, we state the following

LEMMA. *Let T be a sufficient statistic for $\theta \in \Omega$, and let G be a group of 1:1 transformations g on X leaving Ω invariant. Then if $\varphi(x)$ is invariant under G, $E[\varphi(X) \mid t]$ is almost invariant under G.*

We can now state the desired comparison in the following

THEOREM 7.1. *Let G be a group of 1:1 transformations on X, let $\bar{G}$ be the induced group of transformations on Ω, let $v(\theta)$ be maximal invariant under $\bar{G}$, and suppose that $\bar{G}$ leaves ω and $\Omega - \omega$ invariant. Suppose further that T is a sufficient statistic for Ω, and that $\{P_\theta^T\}$, $\theta \in \Omega$, is boundedly complete. Then a necessary and sufficient condition that the power of a test $\psi(T)$ be a function only of $v(\theta)$, is that $\psi(t)$ be almost invariant under G.*

This theorem is an immediate extension of some results of Wolfowitz [35].

Theorem 7.1 together with the results of section 5 proves that the standard tests of the general linear hypothesis, Hotelling's T^2-problem and the hypothesis concerning the multiple correlation coefficient possess the optimum property that was obtained for these problems by Hsu and Simaika, respectively. The method of proof indicated here is due to Wolfowitz [35].

8. Most stringent tests. We shall now turn to the third aspect of the theory: Optimum properties defined with reference to the whole class of alternatives, and attainable with no restrictions imposed on the class of tests. In the present section we shall consider the property of stringency. Wald [25] defines a test φ to be most stringent if it minimizes

$$\sup_{\theta \in \Omega - \omega} [\beta_\alpha^*(\theta) - \beta_\varphi(\theta)], \tag{8.1}$$

where β_α^* again denotes the envelope power, and β_φ the power of φ. The rationale of this definition is clear. The difference $\beta_\alpha^*(\theta) - \beta_\varphi(\theta)$ measures the amount by which the test falls short at the alternative θ of the power that could be attained against this particular alternative. A test φ is therefore most stringent if it minimizes its maximum shortcoming.

A theory of most stringent tests was developed by Hunt and Stein [5], who based it on the notion of invariance. Consider, as in section 5, a group G of measurable 1:1 transformations on $\mathfrak{X}$ leaving the problem invariant. Hunt and Stein obtained their results in connection with the following groups of transformations.

(i) $gx = x + c$, $-\infty < c < \infty$, x a real variable;
(ii) $gx = ax$, $0 < a$, x a real variable;
(iii) $gx = ax + c$, $0 < a$, $-\infty < c < \infty$, x a real variable;
(iv) the group of orthogonal transformations on a Euclidean space;
(v) any finite group.

THEOREM 8.1. (Hunt and Stein). *If G is the direct product of a finite number of groups of types* (i)–(v), *and if G leaves the problem invariant, that is, if G satisfies* (5.4), *then there exists a most stringent test invariant under G.*

Actually, it is not necessary here to require that G be a direct product. The result holds also if the factoring of G is according to normal subgroups, where the normal subgroup at each stage and the final factor group are of the types mentioned. In the light of this one may omit type (iii) from the list since it has a normal subgroup of type (i) with factor group of type (ii).

The proof of Theorem 8.1 is based on the following lemma, which has applications to many related problems.

LEMMA (Hunt and Stein). *If G is a direct product of a finite number of groups of types* (i)–(v) *then given any function f over $\mathfrak{X}$ $(0 \leq f(x) \leq 1)$ there exists a function F $(0 \leq F(x) \leq 1)$ such that F is invariant under $\hat{G}$, and*

$$(8.2) \qquad \inf_{g \in G} \int f(gx)\varphi(x)\, d\mu(x) \leq \int F(x)\varphi(x)\, d\mu(x) \leq \sup_{g \in G} \int f(gx)\varphi(x)\, d\mu(x)$$

for all φ that are integrable μ.

It follows from Theorem 8.1 that if there exists a uniformly most powerful invariant test, this test is most stringent. In this way Hunt and Stein show, for example, (see in this connection section 5), that the likelihood ratio test of the general univariate linear hypothesis is most stringent. A question that is left open is the uniqueness of such a most stringent test.

In general, the possibility therefore remains that there might exist another most stringent test uniformly more powerful than the invariant one. In certain particular cases this possibility can be ruled out by the following consideration. Suppose that Ω is a subset of a Euclidean space and that every point of ω is a limit point of $\Omega - \omega$. Suppose further that for any test ϕ, $E_\theta\phi(X)$ is continuous in θ. Then clearly, if ϕ_1 is similar of size α for testing ω and ϕ_2 is of size $\leqq \alpha$ but not similar, ϕ_2 can not be uniformly as powerful as ϕ_1. Hence any test that is admissible among all similar tests of size α is also admissible among the totality of tests of size $\leqq \alpha$. Now admissibility among all similar tests is sometimes not too difficult to prove. For the likelihood ratio test of the general linear univariate hypothesis, for example, it is an immediate consequence of the properties of this test proved by Wald [23] and Hsu [4].

The following alternative method for obtaining most stringent tests is also mentioned by Hunt and Stein.

THEOREM 8.2. (Hunt and Stein). *Let $\Omega - \omega$ be partitioned into disjoint subsets Ω_δ such that $\beta^*_\alpha(\theta)$ is constant on each Ω_δ, and let φ_δ be the test that maximizes $\inf_{\theta \in \Omega_\delta} \beta_{\varphi_\delta}(\theta)$. Then if $\varphi_\delta = \varphi$ is independent of δ, φ is most stringent.*

This result may be supplemented by the following method for finding tests that maximize $\inf_{\theta \in \omega_1} \beta_\varphi(\theta)$ over a given set of alternatives ω_1 (not necessarily satisfying the conditions imposed above on the Ω_δ's).

THEOREM 8.3. *Suppose additive classes of sets have been defined over* ω *and* ω_1, *and consider probability measures* λ *and* λ_1 *over* ω *and* ω_1. *Let the functions* $f_\theta(x)$ *be generalized probability densities with respect to* μ, *so that* $h(x) = \int_\omega f_\theta(x)\, d\lambda(\theta)$ *and* $h_1(x) = \int_{\omega_1} f_\theta(x)\, d\lambda_1(\theta)$ *are again probability densities with respect to* μ. *Let* φ *be the most powerful test of size* α *for testing the simple hypothesis* $H:h$ *against the simple alternative* h_1, *and suppose that the power of* φ *against* h_1 *is* β. *Then if*

$$\begin{aligned} E_\theta\varphi(x) &\leq \alpha \quad \text{for all} \quad \theta \in \omega, \\ E_\theta\varphi(x) &\geq \beta \quad \text{for all} \quad \theta \in \omega_1, \end{aligned} \tag{8.3}$$

φ *maximizes* $\inf_{\theta \in \omega_1} \beta_\varphi(\theta)$ *at level of significance* α.

This method, when applicable, has the advantage of giving the totality of most stringent tests (see in this connection Theorem 3.1) and hence of settling the question of admissibility. However, in many applications probability measures λ, λ_1 with the desired properties do not exist but instead only sequences $\lambda^{(n)}$, $\lambda_1^{(n)}$, which satisfy the conditions in the limit. In this case again only the weak conclusion is possible: The test obtained is most stringent but has not been proved admissible. (For an example in which the analogous method has been carried through in detail for an estimation problem, see [22]).

Actually, the two methods are closely related, as can be seen from the proof of the main lemma. In those cases in which there exists a group G giving the maximum possible reduction, the group $\bar{G}$ induces a partition of Ω (through the equivalence: $\theta_1 \sim \theta_2$ if there exists $\bar{g}$ such that $\theta_2 = \bar{g}\theta_1$), just into ω and the sets Ω_δ. (This is so mainly because, as was shown by Hunt and Stein, the envelope power remains invariant under any transformations that leave the problem invariant.) Then the measures λ, λ_δ over ω, Ω_δ respectively, which figure in the application of Theorems 8.2 and 8.3, become invariant measures over $\bar{G}$ through the obvious 1:1 mapping from ω and the Ω_δ's respectively to $\bar{G}$. Thus the second method will allow the strong conclusion when the group $\bar{G}$ involved in the first method possesses a finite invariant measure [types (iv) and (v)] but not if any of its factors are of type (i)–(iii).

To conclude this section we shall give an example where the method of invariance leads only to a partial reduction but where the solution may be completed by certain additional considerations. Suppose that $(X_1, \cdots, X_n)$ is a sample from a normal distribution with mean ξ and variance σ^2, both unknown, and that we wish to find the most stringent test of the hypothesis H: $\sigma = 1$ against the alternatives $\sigma \neq 1$. Theorem 8.1 reduces the problem to the statistic $Y = \Sigma(X_i - \bar{X})^2$, but among the tests of H based on this statistic there does not exist a uniformly most powerful one. It may also be shown [8] that no further reduction is possible by means of the method of invariance.

However, one may now consider the problem of finding the most stringent test based on Y. (The envelope power function $\beta^*(\xi, \sigma)$ that must be used naturally is not the one for Y but that for the original problem.) From an argument given in [6] it follows that this test is of the form

$$\varphi_{k_1,k_2}\text{: reject when } Y < k_1 \text{ or } > k_2 ,$$

where k_1, k_2 are determined by the two conditions

(i) $P(\text{rejection} \mid \sigma = 1) = \alpha$,

(ii) $\sup_{\sigma<1} [\beta^*_\alpha(\xi, \sigma) - \beta_{\varphi_{k_1,k_2}}(\sigma)] = \sup_{\sigma>1} [\beta^*_\alpha(\xi, \sigma) - \beta_{\varphi_{k_1,k_2}}(\sigma)]$.

Here $\beta^*_\alpha(\xi, \sigma)$ is independent of ξ and can be obtained from a table of the χ^2-distribution (with n degrees of freedom for $\sigma < 1$ and n-1 degrees of freedom for $\sigma > 1$ as can be seen from (3.6)). Hence k_1 and k_2 can be computed fairly easily.

Another problem that may be treated in this way is the hypothesis of equality of variances for two normal samples. If the two samples are of equal size, there exists a uniformly most powerful invariant test for a suitable group of transformations. However, if the sample sizes are different the method of invariance reduces the problem only to $\Sigma(X_i - \bar{X})^2 / \Sigma(Y_i - \bar{Y})^2$, and the cut off points giving the most stringent test may be determined by an argument analogous to that given above.

This method may be extended to allow determination of most stringent test of hypotheses such as H: $\sigma_1 \leqq \sigma \leqq \sigma_2$. This requires a certain modification of Theorem 1 of [6], which is easily obtained. One finds agains that one may restrict consideration to a one-parameter family of tests (determined by a somewhat different condition than above), and that among these the most stringent test is obtained by the analogue of condition (ii) above.

If should be mentioned that the results of [6] apply also to the hypothesis specifying the value of the parameter in a binomial or Poisson distribution. This is easily seen since in either case the distributions of Ω are absolutely continuous with respect to a common sigma finite measure and since for the appropriate choice of this measure the generalised density is of the form assumed for the density in [6]. Hence in both the binomial and the Poisson case the most stringent test is determined by conditions analogous to (i) and (ii) above.

9. Tests that minimize the maximum loss. In the Neyman-Pearson theory one classifies the errors into two kinds: Rejecting the hypothesis when it is true, accepting it when it is false. One may however analyze the situation further and distinguish, say, between accepting when one or some other alternative is true. Thus one is led to introduce the losses that result in a given situation from the various possible errors, and to look for a test that, in an appropriate sense, minimizes the expected loss. This possibility was mentioned by Neyman and Pearson [17], and was taken as the starting point of his general theory by Wald (see for example [24]).

In order to stay within the framework of this exposition we shall here introduce losses only for the errors of accepting the hypothesis when it is false, while still demanding that the probability of rejection when the hypothesis is true should not exceed α. Actually, there are many cases where this seems to be a reasonable formulation. For it frequently happens that the two types of error entail consequences of such completely different nature that the resulting losses cannot be measured on a common scale while usually the different errors of the same type are comparable.

We shall therefore assume that for each $\theta \epsilon \Omega - \omega$ there is defined a $W(\theta)$, which measures the loss resulting from acceptance of H when θ is true. The risk which one runs by using a test φ, when $\theta \epsilon \Omega - \omega$ is the true parameter value is given by the expected loss $R_\varphi(\theta) = W(\theta) E_\theta[1 - \varphi(X)]$. When a uniformly most powerful test exists for the hypothesis in question, this test also minimizes the expected loss uniformly for θ in $\Omega - \omega$. In the contrary case one may again restrict the class of tests in some way, so that within the restricted class there exists a uniformly most powerful test, and hence a test that uniformly minimizes the expected loss. Alternatively we may again consider some optimum property of the risk function $R_\varphi(\theta)$ as a whole. We shall here consider the minimax principle introduced by Wald, and seek a test, which, subject to $E_\theta\varphi(X) \leq \alpha$ for all $\theta \epsilon \omega$, minimizes

$$\sup_{\theta \epsilon \Omega - \omega} W(\theta) E_\theta[1 - \varphi(X)],$$

the maximum risk.

If one introduces losses also for the other type of error it is easy to see that for a suitably chosen loss function the definition of minimax expected loss coincides with that of stringency. It is therefore not surprising that the methods of the previous section can be extended to cover the problems considered in the present one. (They are actually much more general, and may be applied also, for example, to the problem of point estimation, and in fact to the general decision problem).

From the lemma of Hunt and Stein stated in the previous section we immediately obtain the following extension of Theorem 8.1.

THEOREM 9.1. *If G is a group of transformations leaving the hypothesis and the class of alternatives invariant, if G can be factored by normal subgroups into factors of types* (i)–(v), *and if the loss function $W(\theta)$ is invariant under $\bar{G}$, then there exists a test φ invariant under G and minimizing*

$$\sup_{\theta \epsilon \Omega - \omega} W(\theta) E_\theta[1 - \varphi(X)]. \tag{9.1}$$

It follows that when a uniformly most powerful invariant test exists, this test has the property of minimizing the maximum expected loss with respect to any invariant loss function. Thus Student's test, for example, minimizes the maximum risk for any loss function that depends only on $|\xi|/\sigma$.

Clearly the second method mentioned in section 8 can be extended in an analogous manner if in Theorem 8.2 one replaces the sets Ω_δ by sets over which $W(\theta)$ is constant.

Again it may happen that the method of invariance does not reduce the problem sufficiently far but that the solution may be completed by other considerations. Let us once more consider the hypothesis $H: \sigma = 1$ of the previous section, and let us suppose that the loss function has the necessary invariance property, so that it is a function only of σ but not of the unknown mean. It follows from Theorem 9.1 that there exists a test minimizing the maximum risk, which is a function only of $Y = \Sigma(X_i - \bar{X})^2$. From [6] it is easily seen that a test φ_{k_1,k_2} which rejects when $Y < k_1$ or $> k_2$, has the desired property if its size is α and if in addition

$$\sup_{\sigma<1} W(\sigma)E_\sigma[1 - \varphi(Y)] = \sup_{\sigma>1} W(\sigma)E_\sigma[1 - \varphi(Y)]. \tag{9.2}$$

It follows that depending on the choice of $W(\sigma)$ the solution may be any member of the one-parameter family of tests φ_{k_1,k_2} of size α.

Under the conditions of Theorem 9.1, when a uniformly most powerful invariant test exists, this also maximizes the average power for a large class of weight functions. If there exists a common finite invariant measure over the sets Ω_δ in the sense indicated in section 8, the uniformly most powerful invariant test will maximize the average power with this measure as weight function, over Ω_δ for all δ. It follows that it maximizes the average power over $\Omega - \omega$ with respect to any weight function for which the conditional distribution over each Ω_δ is the above invariant measure. If the invariant measure over the Ω_δ's is not finite one can obtain analogous results with respect to a sequence of weight functions invariant in the limit. The results indicated here are much weaker than those obtained for the general linear univariate hypothesis by Wald [23] and Hsu [4] under the restriction to similar regions. However their results are no longer valid when this restriction is omitted.

10. Applications to sequential analysis. So far we have restricted consideration to the case that the hypothesis is to be tested on the basis of a preassigned experiment. However, frequently there is available for this purpose a large class of experiments, and the selection of an optimum experiment out of this class is part of the problem. We shall consider here only the following situation, which has recently been studied extensively (see Wald [28, 29]). There is given a sequence of random variables $X_1, X_2, \cdots$ whose joint distribution is known to belong to some family $\mathfrak{F} = \{P_\theta\}$, $\theta \in \Omega$; the hypothesis specifies some subfamily: $\theta \in \omega$. The X's are observed one by one, and the decision, whether or not to continue experimentation at any given stage, is allowed to depend on the observations taken up to that point. Thus the number n of observations that will be taken is a random variable whose distribution depends on θ. Usually, by an appropriate choice of stopping rule, there may be effected a considerable saving in the expectation of the number of observations necessary to achieve a given discrimination between hypothesis and alternatives. The problem is to determine the stopping rule and test that minimizes this expectation.

As we have seen in the previous sections the principal methods for obtaining optimum tests consist in reducing the problem to that of testing a simple hypothesis against a simple alternative. This basic problem was solved in the non-sequential case by Neyman and Pearson (Theorem 3.1). The solution of the much more difficult corresponding sequential problem was obtained for a large class of cases by Wald and Wolfowitz [31] in the following

THEOREM 10.1. *Let X_1, X_2, $\cdots$ be identically and independently distributed. It is desired to test the hypothesis that the common probability density of the X's is $f(x)$ against the alternative that it is $g(x)$. Given two numbers $0 < \alpha < \beta < 1$, there exists a test which, subject to the condition*

$$\begin{aligned} &P\,(\text{rejection} \mid f) \leq \alpha \\ &P\,(\text{rejection} \mid g) \geq \beta, \end{aligned} \tag{10.1}$$

minimizes simultaneously $E_f(n)$ and $E_g(n)$, the expected number of observations computed for the distributions f and g. This test is given in terms of two numbers A and B by the following rule. After m observations have been taken,

$$\textit{reject if } \frac{g(x_1) \cdots g(x_m)}{f(x_1) \cdots f(x_m)} > A,$$

$$\textit{accept if } \frac{g(x_1) \cdots g(x_m)}{f(x_1) \cdots f(x_m)} < B,$$

$$\textit{take another observation if } B < \frac{g(x_1) \cdots g(x_m)}{f(x_1) \cdots f(x_m)} < A.$$

Here A and B are determined so that condition (10.1) *holds with the inequality signs replaced by equality.*

So as to be able to treat the various problems considered non-sequentially in the previous sections one would have to extend this theorem at least to the case that the variables X_1, X_2, $\cdots$ form a set of equivalent variables in the sense of de Finetti [1]. Instead, we shall here restrict ourselves to a few problems that can be solved on the basis of Theorem 10.1. All of the tests discussed below were derived from various points of view and some of their properties were discussed by Girshick in his important "Contributions to the theory of sequential analysis". *Annals of Math. Stat.*, vol. 17 (1946) pp. 123–143 and 282–298, and by Wald in his basic book on the subject [28].

It is convenient here to modify slightly the formulation of the problem of hypothesis testing. Let the parameter space Ω be divided into three sets, the set ω_0 specified by the hypothesis, the class of alternatives ω_1, and a region of indifference $\Omega - \omega_0 - \omega_1$ where we do not much care whether the hypothesis is accepted or rejected (see [28]). Let us denote the sequential random variable $(X_1, \cdots, X_n)$ by X. Then we wish to determine a sequential test φ, which, subject to

$$\begin{aligned} &E_\theta\varphi(X) \leq \alpha \text{ for } \theta \in \omega_0 \\ &E_\theta\varphi(X) \geq \beta \text{ for } \theta \in \omega_1, \end{aligned} \tag{10.2}$$

minimizes $\sup_{\theta \epsilon \omega_0+\omega_1} E_\theta(n)$. (Actually, this is a rather artificial formulation. The natural requirement is the minimization of $\sup_{\theta \epsilon \Omega} E_\theta(n)$ but this is a much more difficult problem.) The reduction to the problem of testing a simple hypothesis against a simple alternative is achieved by the following obvious extension of Theorem 8.3.

THEOREM 10.2. *Let* λ_0, λ_1 *be distributions over* ω_0, ω_1 *respectively, and let* φ *be a test, which subject to*

$$
\begin{aligned}
&\int_{\omega_0} E_\theta\varphi(X)d\lambda_0(\theta) \leq \alpha \\
&\int_{\omega_1} E_\theta\varphi(X)d\lambda_1(\theta) \geq \beta,
\end{aligned}
\tag{10.3}
$$

minimizes $\sup_{i\epsilon\{0,1\}} \int E_\theta(n)\, d\lambda_i(\theta)$. *Then if* φ *satisfies* (10.2) *and*

$$E_\theta(n) \leq \sup_{i\epsilon\{0,1\}} \int E_\theta(n)\, d\lambda_i(\theta) \text{ for all } \theta \,\epsilon\, \omega_0 + \omega_1, \tag{10.4}$$

φ *minimizes* $\sup_{\omega_1+\omega_2} E_\theta(n)$ *subject to* (10.2).

As in section 3 we can make certain trivial applications to problems concerning a single real parameter such as testing the hypothesis $H\colon p \leq p_0$ against the alternatives $p \geq p_1$ ($p_0 < p_1$), where p is the probability of success in a binomial sequence of trials. In this example condition (10.2) of Theorem 10.2 obviously is satisfied when λ_0 and λ_1 assign probability 1 to p_0 and p_1 respectively. Hence the probability ratio test for testing $p = p_0$ against $p = p_1$ has the desired properties, whenever (10.4) holds, that is, whenever $E_p(n)$ attains its maximum between p_0 and p_1.

The following is another example that may be solved in this manner. Let $X_1, X_2, \cdots ; Y_1, Y_2, \cdots$ be independently normally distributed, all with unit variance and means $E(X_i) = \xi$, $E(Y_i) = \eta$. In order to test the hypothesis $H\colon \xi \geq \eta$ against the alternatives $\eta - \xi \geq \delta$ where $\delta > 0$ is given, a pair (X_1, Y_1) is observed. If after this observation experimentation continues another pair (X_2, Y_2) is observed, etc. In this case we may take for λ_0, λ_1 the distributions that assign probability 1 to the parameter points $(\xi, \eta) = (0, 0)$ and $\left(-\frac{\delta}{2}, \frac{\delta}{2}\right)$ respectively. Then the probability ratio after m observations is given by

$$\frac{\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\left(x_i+\frac{\delta}{2}\right)^2-\frac{1}{2}\sum_{i=1}^{m}\left(y_i-\frac{\delta}{2}\right)^2\right]}{e^{-\frac{1}{2}\Sigma x_i^2-\frac{1}{2}\Sigma y_i^2}} = e^{-(m\delta^2/4)+\delta[\Sigma y_i-\Sigma x_i]}. \tag{10.5}$$

Since the distribution of $Y - X$ depends only on $\eta - \xi$, it is easily seen that condition (10.2) is satisfied.

Some further results can be obtained through extension to the sequential case of Theorems 8.1 and 9.1.

THEOREM 10.3. *Suppose that G is of the type described in Theorem 9.1, let $Y = f(X_1, X_2, \cdots)$ be maximal invariant under G, let $v(\theta)$ be maximal invariant under $\bar{G}$, and let the set of values of $v(\theta)$ corresponding to ω_0 and ω_1 be $\bar{\omega}_0$ and $\bar{\omega}_1$, respectively. If among all tests of $\bar{\omega}_0$ against $\bar{\omega}_1$ based on Y, the test φ minimizes* $\sup_{v(\theta)\epsilon\bar{\omega}_0+\bar{\omega}_1}$ - $E_\theta(n)$ *subject to*

$$E_\theta\varphi(Y) \leq \alpha \text{ if } v(\theta)\epsilon\bar{\omega}_0$$
$$E_\theta\varphi(Y) \geq \beta \text{ if } v(\theta)\epsilon\bar{\omega}_1, \tag{10.6}$$

then φ also minimizes $\sup_{\omega_0+\omega_1} E_\theta(n)$ *among all tests based on the X's and which satisfy* (10.2).

As an example consider the problem of testing the hypothesis $\sigma \leq \sigma_0$ against the alternatives $\sigma \geq \sigma_1$ $(\sigma_0 < \sigma_1)$ when the X's are identically, independently normally distributed with unknown mean and variance. Since the problem remains invariant under a common translation of the X's we can take for Y of the theorem $Y = (X_2 - X_1, X_3 - X_1, \cdots)$. Equivalently we may take as our new sequence of variables $(Y_1, Y_2, \cdots)$ where

$$Y_k = \frac{kX_{k+1} - (X_1 + \cdots + X_k)}{\sqrt{k(k+1)}}. \tag{10.7}$$

Then $Y_1, Y_2, \cdots$ are independently normally distributed with zero mean and the same variance as the X's. Hence the problem reduces to a type which we have already considered. The optimum test is based on

$$\sum_{i=1}^{m} Y_i^2 = \sum_{i=1}^{m+1}\left(X_i - \frac{X_1 + \cdots + X_{m+1}}{m+1}\right)^2.$$

It may be worth pointing out that Theorems 3.2, 8.3, 10.2 all are special cases of simple results in the general theory of statistical decision functions, of which the following is the prototype. (For a detailed treatment of this theory see, for example, [30]). Let $\{P_\theta\}$, $\theta\epsilon\Omega$, be the family of possible distributions of a random variable X, and let $\{\delta\}$ be a family of decision functions. The loss resulting from the use of $\delta(x)$ when P_θ is the true distribution is $W[\theta, \delta(x)]$ and the risk function associated with δ is $R_\delta(\theta) = E_\theta W[\theta, \delta(X)]$. Let λ be a probability measure over Ω, and let δ_λ be a decision function that minimizes $\int R_\delta(\theta)\, d\lambda(\theta)$. Then if λ is such that

$$R_{\delta_\lambda}(\theta) \leqq \int R_{\delta_\lambda}(\zeta)\, d\lambda(\zeta) \text{ for all } \theta \epsilon \Omega, \tag{10.8}$$

δ_λ minimizes $\sup_\theta R_\delta(\theta)$.

PROOF. Let δ^* be any other decision function. Then

$$\sup_\theta R_{\delta_\lambda}(\theta) \leq \int R_{\delta_\lambda}(\theta)\, d\lambda(\theta) \leqq \int R_{\delta^*}(\theta)\, d\lambda(\theta) \leqq \sup_\theta R_{\delta^*}(\theta).$$

In an analogous manner one can give an extension of Theorems 8.1, 9.1, 10.3.

11. Two sided tests considered as 3-decision problems. In a number of important special problems the hypothesis specifies the value of a real valued parameter or states that this parameter lies in a certain interval, and it is desired to test this hypothesis against the obvious two-sided class of alternatives. It seems that in nearly any problem of this kind that would arise in practice one would want to decide when rejecting the hypothesis, whether the true parameter value lies below or above the hypothetical ones. If for example one rejects the hypothesis that the means of two normal populations are equal, one usually wants to decide which of the two is larger. It would therefore seem most natural to formulate such problems as 3-decision problems.

Problems of this kind, as all problems of hypothesis testing, naturally are special cases of the general decision problem formulated by Wald. We shall here consider the case that upper bounds are given for the probabilities of certain types of errors and thereby obtain a formulation, which is closely analogous to the classical formulation of hypothesis testing discussed in this paper, and which will allow immediate application of a large portion of the theory discussed here.

Consider the case that Ω is partitioned into 3 parts, ω, ω_1, ω_2 where in a certain sense ω lies between ω_1 and ω_2. We wish to test the hypothesis $H: \theta \epsilon \omega$. When we reject the hypothesis, we shall reach either decision D_1 that $\theta \epsilon \omega_1$ or decision D_2 that $\theta \epsilon \omega_2$. Correspondingly we prescribe two positive numbers α_1, α_2 and impose the restriction that

$$P_\theta(D_1) \leq \alpha_1 \text{ if } \theta \epsilon \omega + \omega_2$$
$$P_\theta(D_2) \leq \alpha_2 \text{ if } \theta \epsilon \omega + \omega_1. \tag{11.1}$$

Subject to this condition it is desired to maximize

$$P_\theta(D_1) \text{ for } \theta \epsilon \omega_1$$
$$P_\theta(D_2) \text{ for } \theta \epsilon \omega_2. \tag{11.2}$$

A test will now consist of two non-negative functions ϕ_1 and ϕ_2 satisfying

$$\phi_1(x) + \phi_2(x) \leq 1, \tag{11.3}$$

with the convention that when $X = x$ the decision D_i will be taken with probability $\phi_i(x)$ $(i = 1, 2)$.

There is no difficulty concerning the extension of the notions of invariance or sufficient statistic, in fact these notions obviously apply to the general decision problem. The notion of unbiasedness is extended in the obvious way by the condition

$$P_\theta(D_1) \geq \alpha_1 \text{ for } \theta \epsilon \omega_1$$
$$P_\theta(D_2) \geq \alpha_2 \text{ for } \theta \epsilon \omega_2. \tag{11.4}$$

One then obtains the following

THEOREM 11.1. *Suppose that for testing the hypothesis H_1: $\theta \epsilon \omega + \omega_2$ against the alternatives $\theta \epsilon \omega_1$ at level of significance α_1, the test ϕ_1 among all unbiased tests* *is uniformly most powerful in* $\omega + \omega_2$ *and uniformly least powerful in* ω_1, *and that* ϕ_2 *has the analogous property for testing* $H_2 : \theta \epsilon \omega + \omega_1$ *against* $\theta \epsilon \omega_2$ *at significance level* α_2. *If* $\phi_1(x) + \phi_2(x) \leqq 1$ *for all* x, *then among all procedures satisfying* (11.1) *and* (11.4), *the procedure* (ϕ_1, ϕ_2) *uniformly maximizes the probability of a correct decision*. (If the tests ϕ_1, ϕ_2 take on only the values 0 and 1, the condition $\phi_1(x) + \phi_2(x) \leqq 1$ states that the rejection region of each of the two hypotheses is contained in the acceptance region of the other.)

As an example consider the case that $X_1, \cdots, X_n$ are independently, normally distributed with common mean ξ and variance σ^2. Suppose we wish to test the hypothesis that $\sigma_1 \leqq \sigma \leqq \sigma_2$ where σ_1 may equal σ_2. Then it follows from Theorem 11.1 that among all unbiased procedures of level (α_1, α_2), there exists one that maximizes the probability of a correct decision uniformly in ξ, σ. This is the procedure under which decision D_1 or D_2 is taken as $\Sigma(x_1 - \bar{x})^2 \leqq k_1$ or $\geqq k_2$ and the hypothesis is accepted otherwise. Here the k's are determined by

$$P(\Sigma(x_i - \bar{x})^2 \leqq k_1 \mid \sigma_1) = \alpha_1 \tag{11.5}$$
$$P(\Sigma(x_i - \bar{x})^2 \geqq k_2 \mid \sigma_2) = \alpha_2 .$$

REFERENCES

[1] B. DE FINETTI, "La prévision: ses lois logiques, ses sources subjectives," *Annales de l'Institut Henri Poincaré*, Vol. 7 (1937), p. 1.

[2] P. R. HALMOS AND L. J. SAVAGE, "Applications of the Radon-Nikodym theorem to the theory of sufficient statistics," *Annals of Math. Stat.*, Vol. 20 (1949), p. 225.

[3] P. L. HSU, "Analysis of variance from the power function standpoint," *Biometrika*, Vol. 32 (1941), p. 62.

[4] P. L. HSU, "On the power function of the E^2-test and the T^2-test," *Annals of Math. Stat.*, Vol. 16 (1945), p. 278.

[5] G. HUNT AND C. STEIN, "Most stringent tests of statistical hypotheses," unpublished.

[6] E. L. LEHMANN, "On families of admissible tests," *Annals of Math. Stat.*, Vol. 18 (1947), p. 97.

[7] E. L. LEHMANN, "On the theory of testing composite hypotheses with one constraint," *Annals of Math. Stat.*, Vol. 18 (1947), p. 473.

[8] E. L. LEHMANN, "On the totality of transformations leaving a statistical problem invariant," unpublished.

[9] E. L. LEHMANN AND H. SCHEFFÉ, "On the problem of similar regions," *Proc. Nat. Acad. Sci.*, Vol. 33 (1947), p. 382.

[10] E. L. LEHMANN AND H. SCHEFFÉ, "Completeness, similar regions and unbiased estimation," unpublished.

[11] E. L. LEHMANN AND C. STEIN, "Most powerful tests of composite hypotheses. I. Normal distributions," *Annals of Math. Stat.*, Vol. 19 (1948), p. 495.

[12] E. L. LEHMANN AND C. STEIN, "On the theory of some non-parametric hypotheses," *Annals of Math. Stat.*, Vol. 20 (1949), p. 28.

[13] J. NEYMAN, "Su un teorema concernente le cosidette statistiche sufficienti," *Giorn. Ist. Ital. Att.*, Vol. 6 (1935), p. 320.

[14] J. NEYMAN, "Sur la vérification des hypothèses statistiques composées," *Bull. Soc. Math. France*, Vol. 63 (1935), p. 1.

[15] J. NEYMAN, "Outline of a theory of statistical estimation based on the classical theory of probability," *Phil. Trans. Roy. Soc. London*, Series A, Vol. 236 (1937), p. 333.

[16] J. NEYMAN AND E. S. PEARSON, "On the use and interpretation of certain test criteria for purposes of statistical inference," *Biometrika*, Vol. 20A (1928), p. 175 and p. 263.

[17] J. NEYMAN AND E. S. PEARSON, "On the testing of statistical hypotheses in relation to probability a priori," *Proc. Camb. Phil. Soc.*, Vol. 29 (1933), p. 492.

[18] J. NEYMAN AND E. S. PEARSON, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. Roy. Soc. London*, Series A, Vol. 231 (1933), p. 289.

[19] J. NEYMAN AND E. S. PEARSON, "Contributions to the theory of testing statistical hypotheses. I. Unbiased critical regions of type A and type A_1," *Stat. Res. Mem.*, Vol. 1 (1936), p. 1.

[20] E. J. G. PITMAN, "Tests of hypotheses concerning location and scale parameters," *Biometrika*, Vol. 31 (1939), p. 200.

[21] J. B. SIMAIKA, "On an optimum property of two important statistical tests," *Biometrika*, Vol. 32 (1941), p. 70.

[22] C. STEIN AND A. WALD, "Sequential confidence intervals for the mean of a normal distribution with known variance," *Annals of Math. Stat.*, Vol. 18 (1947), p. 427.

[23] A. WALD, "On the power function of the analysis of variance test," *Annals of Math. Stat.*, Vol. 13 (1942), p. 434.

[24] A. WALD, "On the principles of statistical inference," *Notre Dame Math. Lectures*, No. 1 (1942).

[25] A. WALD, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Am. Math. Soc.*, Vol. 54 (1943), p. 426.

[26] A. WALD, "Statistical decision functions which minimize the maximum risk," *Annals of Math.*, Vol. 46 (1945), p. 265.

[27] A. WALD, "An essentially complete class of admissible decision functions," *Annals of Math. Stat.*, Vol. 18 (1947), p. 549.

[28] A. WALD, *Sequential analysis*, John Wiley and Sons, 1947.

[29] A. WALD, "Foundations of a general theory of sequential decision functions," *Econometrica*, Vol. 15 (1947), p. 279.

[30] A. WALD, "Statistical decision functions," *Annals of Math. Stat.*, Vol. 20 (1949), p. 165.

[31] A. WALD AND J. WOLFOWITZ, "Optimum character of the sequential probability ratio test," *Annals of Math. Stat.*, Vol. 19 (1948), p. 326.

[32] S. S. WILKS, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Annals of Math. Stat.*, Vol. 9 (1938), p. 60.

[33] J. WOLFOWITZ, "Additive partition functions and a class of statistical hypotheses," *Annals of Math. Stat.*, Vol. 13 (1942), p. 247.

[34] J. WOLFOWITZ, "Non-parametric statistical inference," *Proceedings of the Berkeley Symposium on mathematical statistics and probability* (1949), p. 93.

[35] J. WOLFOWITZ, "The power of the classical tests associated with the normal distribution," *Annals of Math. Stat.*, Vol. 20 (1949), p. 540.

Some related papers not referred to in the text.

[36] T. W. ANDERSON, "On the theory of testing serial correlation," *Skand. Aktuarietidskrift*, (1948), p. 88.

[37] R. A. FISHER, *The design of experiments*, Oliver and Boyd, 1935.

[38] M. N. GHOSH, "On the problem of similar regions," *Sankhya*, Vol. 8 (1948), p. 329.

[39] P. G. HOEL, "Testing the homogeneity of Poisson frequencies," *Annals of Math. Stat.*, Vol. 16 (1945), p. 362.

[40] P. G. HOEL, "Discriminating between binomial distributions," *Annals of Math. Stat.*, Vol. 18 (1947), p. 556.

[41] P. G. HOEL, "On the uniqueness of similar regions," *Annals of Math. Stat.*, Vol. 19 (1948), p. 66.

[42] E. L. LEHMANN, "Some comments on large sample tests," *Proceedings of the Berkeley Symposium on mathematical statistics and probability* (1949), p. 451.

[43] H. B. MANN AND A. WALD, "On the choice of the number of intervals in the application of the χ^2-test," *Annals of Math. Stat.*, Vol. 13 (1942), p. 306.

[44] J. NEYMAN, *Lectures and conferences on mathematical statistics*, Graduate School of the U. S. Dept. of Agriculture, 1938.

[45] J. Neyman, "Basic ideas and some recent results of the theory of testing statistical hypotheses," *Journal Roy. Stat. Soc.*, Vol. cv (1942), p. 292.
[46] J. Neyman, "On a statistical problem arising in routine analysis and in sampling inspection of mass production," *Annals of Math. Stat.*, Vol. 12 (1941), p. 46.
[47] S. N. Roy, "Notes on testing composite hypotheses, I, II," *Sankhya*, Vol. 8 (1947), p. 257 and Vol. 9 (1948), p. 19.
[48] H. Scheffé, "On the theory of testing composite hypotheses with one constraint," *Annals of Math. Stat.*, Vol. 13 (1942), p. 280.
[49] H. Scheffé, "On the ratio of the variances of two normal samples," *Annals of Math. Stat.*, Vol. 13 (1942), p. 371.
[50.] C. Stein, "A two sample test for a linear hypothesis whose power is independent of the variance," *Annals of Math. Stat.*, Vol. 16 (1945), p. 243.
[51] A. Wald, "Asymptotically most powerful tests of statistical hypotheses," *Annals of* Math. Stat., Vol. 12 (1941), p. 1.
[52] A. Wald, "Some examples of asymptotically most powerful tests," *Annals of Math. Stat.*, Vol. 12 (1941), p. 396.
[53] A. Wald, "On the efficient design of statistical investigations," *Annals of Math. Stat.*, Vol. 14 (1943), p. 134.