THE RELATIONSHIP BETWEEN SUFFICIENCY AND INVARIANCE WITH APPLICATIONS IN SEQUENTIAL ANALYSIS

By W. J. Hall[1], R. A. Wijsman[2], and J. K. Ghosh[3]

University of North Carolina, University of Illinois, and Calcutta University

Preface. The main result in this paper relating sufficiency and invariance was originally found by Charles Stein before 1950 but was not published and not widely known. It has since been rediscovered independently by Burkholder in 1958 (reported in [7]), by Hall in 1959 ([18], [19]), and by Ghosh in 1960 [16]; the best theorems of this kind have since been developed by Wijsman. This result is closely related to a theorem of D. R. Cox [9], published in 1952 and widely used in sequential analysis (e.g., [14], [17], [22], [23]) though Cox made no explicit use of invariance concepts. The result, together with extensions (due to Wijsman), related results on transitivity (due to Ghosh), and sequential applications (due to Hall and Ghosh), is now finally published as a joint contribution, with the permission of Stein and Burkholder.

This paper is presented in two parts. Part I, largely written by Hall and Ghosh, discusses the implications of the main result and sketches a proof. It also discusses a result in transitivity and the application of it and the main result to sequential analysis. Several normal theory examples and a sequential rank test are treated in some detail. Part II, largely written by Wijsman, presents the general theory in the subfield mode, including related results on conditional independence and transitivity, and additional examples.

The authors wish to thank H. K. Nandi for the research guidance given to one of them, D. L. Burkholder for helpful discussions, and E. L. Lehmann, J. L. Hodges, Jr. and W. Kruskal for making this joint endeavor possible.

Part I. EXPOSITION AND SEQUENTIAL APPLICATIONS

Received 29 October 1964.

[1] Research was supported in part by the Air Force Office of Scientific Research, the Office of Naval Research, and the National Institutes of Health under Grant GM-10397. Reproduction is permitted for any purposes of the United States Government.

[2] Research was supported by the National Science Foundation under Grants G-11,382 and G-21,507.

[3] Research was supported by a junior research training scholarship grant of the Government of India. Present address: University of Illinois.

Part II. GENERAL THEORY

PART I: EXPOSITION AND SEQUENTIAL APPLICATIONS

I.1. Introduction. We investigate in what sense sufficiency properties are preserved under the invariance principle and thereby obtain an interpretation of the sufficiency of a statistic in the presence of nuisance parameters—an interpretation which facilitates the derivation of some sequential tests of composite hypotheses.

Sufficiency and invariance are *reduction principles*—principles for condensing or reducing the data x to a few statistics which can then be used for purposes of drawing inferences concerning the probability model of the data. Loosely speaking a sufficiency reduction replacing x by $s = s(x)$ discards information which is not relevant to the parameter θ of the model; an invariance reduction (for an explanation of the invariance principle, see [30]) replacing x by $u = u(x)$ discards information about θ which does not pertain solely to a parametric function $\gamma = \gamma(\theta)$ of special interest; the two reductions applied in tandem, replacing x by $v = v(x)$ where $v(x) = v_1(s(x))$ or $v_2(u(x))$, retain only relevant information pertaining solely to γ. It is the interpretation of this latter statement that concerns us here.

More specifically, suppose we consider a family of distributions indexed by θ and some group of transformations on the sample space (e.g., changes in sign, location, scale or order) which leaves the family of distributions unchanged. Then decision procedures will not be affected by the transformations if they are based on invariant functions on the sample space; such invariant functions have distributions depending only on some function, say γ, of θ. We shall show that, loosely speaking, if a statistic s contains all relevant information about θ then a *maximal invariant* function of s contains all the relevant information about γ that is available in any invariant function. We call such functions *invariantly sufficient*; they are sufficient for the family of distributions of a maximal invariant function—or in a sense to be explained, of any invariant function—on the sample space. This result is a consequence of the Stein Theorem presented in this paper. It may be illustrated by the following diagram, in which vertical arrows indicate maximal invariance reductions and horizontal arrows sufficiency reductions:

$$\begin{array}{c|c} \theta & X_\theta \to S_\theta \\ \downarrow & \downarrow \quad\quad \downarrow \\ \gamma & U_\gamma \to V_\gamma \end{array}.$$

Here X_θ may be thought of as the probability model for the data x, with distribution depending on θ; S_θ as the probability model for a sufficient statistic s, whose distribution also depends on θ; U_γ as the probability model for a maximal invariant function u, with distribution depending only on γ; and V_γ as the probability model for an invariantly sufficient statistic v, obtained by either route in the diagram—as a sufficiency reduction on U_γ (definition of invariant sufficiency) or as a maximal invariance reduction on S_θ. In applications it is usually the sufficiency reduction on U_γ that is wanted, while it is the maximal invariance reduction on S_θ that is the easier to perform. That these two routes are equal under certain conditions is the content of the Stein Theorem.

These results seem to be implicitly assumed by other authors. Thus Lehmann [30] derives most of the common significance tests about normal models as uniformly most powerful invariant tests based on these statistics obtained via the upper route ($X_\theta \to S_\theta \to V_\gamma$). In these normal theory problems, sufficiency and invariance frequently reduce the data to a single numerical-valued statistic the sample mean or its magnitude (see I.3), the sample variance (I.3), the t-statistic (I.3), the F-ratio (I.6), Hotelling's T^2-statistic (I.6), the (multiple) correlation coefficient (II.7), and others ([30], [13], [1]). In Lehmann's treatment of rank tests [30], sufficiency and invariance reductions are applied in alternating order. For example, when testing whether two random samples come from the same population with a continuous distribution function against the alternative that one variable is stochastically larger than the other, sufficiency and invariance reductions reduce the two samples to the ranks (in the combined sample) associated with one sample (I.8). Without the Stein Theorem the justification and interpretation of these reductions is not clear.

The situation is similar as regards sequential analysis. Bahadur [4] has shown that reduction by the sufficiency principle is also possible in sequential analysis if the sequence of sufficient statistics, for the data up to stage n, $n = 1, 2, \cdots$, is *transitive* (as defined by him). By interpreting the Stein Theorem as a theorem on conditional independence, it can be shown that the transitivity of a sufficient sequence is preserved under an invariance reduction. Hence, in this case also, one may either reduce first by invariance and then by sufficiency or follow the reverse procedure, and the latter is usually easier.

Moreover, the sufficiency assertion in the Stein Theorem may be used to construct sequential tests for certain kinds of composite hypotheses. The method essentially consists of applying a Wald SPRT for simple hypotheses about γ (or any other sequential test based on likelihood ratios) to what we shall call an *invariantly sufficient sequence* of statistics, the successive values of v. This method has been described by Cox [9] who gave conditions under which the joint density of n terms in this sequence factors conveniently. Application of the Stein Theorem clarifies the motivation of these tests, reinterpreting them through the principle of invariance, and also constitutes a simplification and extension of Cox's factorization theorem. Moreover, it should be noted that Cox's result is imprecisely stated, some vital assumptions having been omitted (see I.5).

In these sequential applications, and also in many nonsequential contexts, invoking invariance can be considered as a way of handling nuisance parameters. (Fraser [12] has considered a generalized sufficiency definition in the presence of nuisance parameters, differing from this approach through invariance. It is shown in [16] that Fraser sufficiency and invariant sufficiency sometimes, though rarely, coincide. See also the end of I.7.) Sometimes we are not concerned with θ as a whole but only some function (say γ) of it—the remaining variability in the parameter space being ascribed to nuisance parameters. For example, a normal distribution with mean θ and known variance may be considered as a two-parameter distribution, one parameter being the magnitude of the mean ($\gamma = |\theta|$) and the other its sign. By invoking invariance under changes in sign, one obtains the magnitude of the sample mean as an invariantly sufficient statistic—a statistic which contains all the information about $|\theta|$ that is available in any invariant statistic, the sign of θ being a nuisance parameter (see I.3). Similarly, with normal mean and variance unknown and $\theta = (\mu, \sigma^2)$, the sample variance is invariantly sufficient for the population variance $\gamma = \sigma^2$ under changes in location, and the t-statistic is invariantly sufficient under changes in scale ($\gamma = \mu/\sigma$); in fact, the t-statistic is invariantly sufficient in a wider context (see II, Example 7.3). Unfortunately, however, invariance theory is not always applicable in problems with nuisance parameters—seldom is it applicable in discrete models.

Invoking invariance in inference or decision problems insures that the error probabilities or risk functions will be independent of the nuisance parameters. However, it may be possible to reduce the error probabilities or the risk functions by using non-invariant decision procedures, and in fact as shown by Stein (see pp. 338–339 in [30]) there are examples where all invariant procedures are inadmissible. On the other hand the (extended) Hunt-Stein Theorem shows that in a number of important cases minimax solutions may be invariant [30], and invariance theory may then provide a useful means of deriving or characterizing minimax procedures.

I.2. Invariantly sufficient statistics. We represent a probability model (space) by $X_\theta = (\mathfrak{X}, \mathfrak{A}, P_\theta)$ where $\mathfrak{X}$ is a sample space of points x, $\mathfrak{A}$ is a given σ-field of subsets of $\mathfrak{X}$, and P_θ is a probability measure on $\mathfrak{A}$. For simplicity, one might like to think of X_θ as a notation for a random sample from a population with density or mass function p_θ. We represent a class of probability models indexed by θ by $X_\Theta = \{X_\theta : \theta \in \Theta\}$ where Θ is some index set.

Any (measurable) function t on $\mathfrak{X}$ induces a new probability model which we denote by $T_\theta = (\mathfrak{I}, \mathfrak{A}^t, P_\theta^t)$ with analogous notations for other functions. Here P_θ^t is the induced probability measure on $\mathfrak{A}^t = t(\mathfrak{A})$ and $\mathfrak{I} = t(\mathfrak{X})$ is the sample space of t (that is, the density or mass function p_θ^t of t is obtained by transformation from p_θ).

We consider a group G with elements g of one-to-one (measurable) transformations from $\mathfrak{X}$ onto itself, and assume, as in [30], that each g induces a transformation $\bar{g}$ from Θ onto itself defined by $P_\theta(gx \in A) = P_{\bar{g}\theta}(x \in A)$, $A \in \mathfrak{A}$, $\theta \in \Theta$.

Hence, the transformed model is among those considered originally. We represent these assumptions symbolically as $gX_\Theta = X_\Theta$, and say *the class of models is invariant under* G. $\bar{G}$ will denote the group with elements $\bar{g}$.

The group G partitions $\mathfrak{X}$ into equivalence classes or orbits. A function t on $\mathfrak{X}$, which is constant on an orbit, is *invariant*. More specifically, t is called *invariant on* $\mathfrak{X}$ *under* G if $t(x) = t(gx)$ for all g and x. If an invariant function u on $\mathfrak{X}$ assumes a different value on each orbit, it is a *maximal invariant*. In other words, u is a *maximal invariant on* $\mathfrak{X}$ *under* G if $u(x) = u(gx)$ for all g and x and if $u(x') = u(x)$ implies the existence of a $g \in G$ for which $x = gx'$. Maximal invariants always exist, and they have the property that all invariant functions are functions of the maximal invariant; also, if t is invariant under G then its distribution depends only on a maximal invariant function, say γ, on Θ under $\bar{G}$ [30]. We denote the probability model corresponding to an invariant statistic t by

$$T_\gamma = (\mathfrak{T}, \mathfrak{A}^t, P_\gamma^t), \qquad \gamma \in \Gamma = \gamma(\Theta).$$

A set $A \in \mathfrak{A}$ is an *invariant set* if $x \in A$ implies $gx \in A$ for every $g \in G$. Any invariant set is of the form $\{x: u(x) \in A^u\}$ where u is a maximal invariant.

A (measurable) function s on $\mathfrak{A}$ is said to be a *sufficient statistic for* X_Θ if for every $A \in \mathfrak{A}$ and $s_0 \in \mathcal{S}$ there is a version of the conditional probability $P_\theta(A \mid s_0) = P_\theta(x \in A \mid s(x) = s_0)$ which does not depend on θ.

If t is any statistic for which $t(gx) = t(gx')$ whenever $t(x) = t(x')$, we say that G *induces a group* G_t *of transformations* g_t *on* $\mathfrak{T}$. Here g_t is defined by $g_t t' = t(gx')$ for $t' \in \mathfrak{T}$ and x' satisfying $t(x') = t'$. Clearly, if u_t is invariant on $\mathfrak{T}$ under G_t, then $u = u_t t (= u_t[t(x)])$ is invariant on $\mathfrak{X}$ under G. Hereafter, we drop the brackets in our notation for composition of functions, writing $z(\cdot) = z_y y(\cdot)$ for $z_y[y(\cdot)]$.

Finally, we shall assume that any sufficient statistic s which we consider is such that G actually induces a group G_s of transformations on the sample space $\mathcal{S}$ of s. Although this assumption holds in all interesting examples known to us, counterexamples can easily be constructed; but without this assumption the invariance reduction on S_6 indicated in the diagram cannot even be defined.

In summary then, our *basic assumptions* are that we are considering a class of probability models X_Θ, a group G of one-to-one transformations on the sample space $\mathfrak{X}$ which leaves the class of models invariant ($gX_\Theta = X_\Theta$), and a sufficient statistic s on $\mathfrak{X}$ which has the property that G induces a group G_s of transformations on the sample space of s.

We now define *invariant sufficiency*, which describes the lower route in the diagram:

Definition. A function v on $\mathfrak{X}$ is invariantly sufficient for X_Θ under G if

(i) v is invariant under G, and

(ii) the conditional probability of any invariant set A given v is parameter-free for $\theta \in \Theta$ (for suitable determination of the conditional probability).

By (i) we may write $v = v_u u$ where v_u is a function on $\mathfrak{U}$ and u is a maximal

invariant on $\mathfrak{X}$. Then a statement equivalent to (ii) is

(ii′) v_u is sufficient for U_Γ,

since $P^u_{\gamma(\theta)}(u \in A^u \mid v_u = v_0)$ is the same as $P_\theta(u(x) \in A^u \mid v(x) = v_0)$. It is tempting to think of (ii) as stating that v is sufficient for the distributions of *any* invariant t but this would require an extended definition of sufficiency which did not require v to be expressible as v_t. (With such a definition, namely that the probability that $t(x)$ lies in any t-set given v is known, the sufficient statistic v may well contain more information about the parameter than does t, but certainly not less.) We can say, however, that v (or rather v_w where $v = v_w w$) is sufficient for the distributions of $w = (v, t)$ where t is any invariant function. It is with these various interpretations in mind that we state loosely that v contains all the information about γ that is available in any invariant function.

We are now prepared to give an informal statement of the main theorem:

STEIN THEOREM. *Under certain assumptions, if s is sufficient for X_Θ and u_s is a maximal invariant on $\mathcal{S}$ under G_s, then $v = u_s s$ is invariantly sufficient for X_Θ under G.*

In the case of discrete distributions (actually, only s need be discrete), the Stein Theorem can be proved without any additional assumptions. The proof is elementary and is given here in two parts:

1.° For any θ and s_0 for which $P_\theta(s(x) = s_0) > 0$, any g, and any invariant set A (i.e., $gA = \{gx: x \in A\} = A$), we have $P_\theta(A \mid s_0) = P_\theta(x \in A, s(x) = s_0)/P_\theta(s(x) = s_0) = P_{\bar{g}\theta}(x \in gA, s(x) = g_s s_0)/P_{\bar{g}\theta}(s(x) = g_s s_0) = P_{\bar{g}\theta}(gA \mid g_s s_0) = P_{\bar{g}\theta}(A \mid g_s s_0)$. Since s is a sufficient statistic, the extreme members are parameter-free so that $P(A \mid s_0) = P(A \mid g_s s_0)$.

2.° Let A be an invariant set and let $P_\theta(s \mid v_0)$ denote $P_\theta(s(x) = s \mid v(x) = v_0)$; then $P_\theta(A \mid v_0) = \sum P_\theta(s \mid v_0) P(A \mid s)$ where the summation is over all s-values for which $u_s(s) = v_0$. Since u_s is a maximal invariant, all those s-values are of the form $g_s s_0$, with s_0 fixed, $g_s \in G_s$. Therefore, $P_\theta(A \mid v_0) = \sum P_\theta(g_s s_0 \mid v_0) \times P(A \mid g_s s_0)$, where the summation is over all g_s. But $P(A \mid g_s s_0) = P(A \mid s_0)$ by 1° so that it factors out of the summation, leaving a sum which is unity. We have thus proved that $P_\theta(A \mid v_0) = P(A \mid s_0)$, free of θ, which concludes the proof.

Part 2° of the proof easily generalizes to the general (instead of discrete) case, by noting from the conclusion of part 1° that $P(A \mid s)$ depends on s only through $u_s s = v$, and writing $P_\theta(A \mid v_0) = E_\theta(P(A \mid s_0) \mid v_0) = P(A \mid s_0)$, which is parameter-free.

Part 1° is not so immediately generalized. In the discrete case, it shows that the conditional probability of an invariant set, given s, is invariant. More generally, the question posed is: Is there a version of the conditional probability which is both parameter-free and invariant? It is only known generally that the conditional probability is *almost invariant* (Lemma 3.1, Part II). Therefore, in order to assure the invariance of $P(A \mid s)$, and with it the Stein Theorem, another assumption has to be made. The choice of the most convenient such assumption depends on the type of problem at hand, and so in Part II three possibilities are

being offered: Assumptions A, B, and C. We indicate below specialized versions of each of these which will enable us to apply the Stein Theorem to a variety of examples; the sufficiency of these assumptions is verified in Part II.

Assumption A is satisfied if every *almost invariant function* on $\mathcal{S}$ is *equivalent to an invariant function*. (We really refer to Assumption A (ii) of Part II since we assume A (i) as part of our *basic assumptions* in Part I.) This is commonly true in parametric problems, a useful sufficient condition being the existence of an invariant measure on G (see pp. 225–228 and 335 in [30] and II.3, II.5). Moreover, it always holds for finite groups, such as sign changes or permutations (see I.3).

Assumption B is, essentially, that there exists an invariant conditional probability distribution, $P(A \mid s) = P(gA \mid g_s s)$, from which the theorem readily follows, and this assumption is easily verified in some important nonparametric applications. In fact, as a useful special case, suppose the sufficient statistic s has the property that any s-set B may be partitioned into sets $B_1, B_2, \cdots$, in such a way that the set of x-values mapping into B_i may be partitioned into i subsets of equal probability (for all θ) on which s is one-to-one; hence, there are a finite number of x-values, all "equally likely", which map into any s-value. Suppose G is any group which induces a group on $\mathcal{S}$. Then $P(A \mid s_0)$ may be taken as the proportion of the x-values mapping into s_0 which are in A. That this $P(A \mid s)$ is invariant is immediate; that it is a version of the conditional probability is straightforward to verify. An example is provided by s being the order statistic(s) corresponding to samples from one or more populations and gx is obtained from x by applying an order-preserving transformation to each observation (see I.8 and Example II.6.1). Another example may be provided by a sufficiency reduction which is simply the dropping of signs in data with a symmetric distribution.

Assumption C concerns regular continuous cases, and the Stein Theorem under this assumption may be considered a rigorous version of the Cox theorem. The conditions are that x has a multivariate (non-singular) continuous distribution, the region of positive density not varying with θ, and the factorization of the joint density of x may be written $g_\theta(s(x))h(x)$ where the transformations g in G, the sufficient statistic s, and the factor h satisfy certain regularity conditions, namely: for all x-values except those lying in an invariant set A_0 having probability zero and satisfying the condition that $s(x) \neq s(x')$ if x, but not x', is in A_0, we have (a) each g is continuously differentiable and both the Jacobian and $h(gx)/h(x)$ depend only on $s(x)$, and (b) s is continuously differentiable with matrix of partial derivatives of maximal rank. Most normal theory examples satisfy these conditions (see I.3 and II.7).

Finally, it is trivial to show that *completeness* of a family of distributions is preserved under invariance reductions. However, in the absence of completeness, *minimality* of sufficiency is not, in general, preserved (see II.3). Hence, it is possible for maximal reductions made by the upper route in the diagram in I.1 to lead to a lesser reduction than maximal reductions made by the lower route.

I.3. Some examples. As a simple illustrative example consider the following: $x = (x_1, x_2)$ where the x_i's are independent and normal with means θ and unit variances, $\Theta = \{\theta: |\theta| < a \text{ (finite or infinite)}\}$, and $G = \{g^+, g^-\}$ where $g^+x = x$ and $g^-x = -x = (-x_1, -x_2)$. Then $\bar{G}$ is found to be $\{\bar{g}^+, \bar{g}^-\}$ where $\bar{g}^\pm\theta = \pm\theta$ so that the class of models is invariant under G. Moreover, $s(x) = x_1 + x_2$ is sufficient for X_Θ, and G induces the group $G_s = \{g_s^+, g_s^-\}$, where $g_s^\pm = \bar{g}^\pm$, on s so that the basic assumptions are satisfied. Moreover, Assumption A holds since G is finite.

The maximal invariants are found to be: $\gamma(\theta) = |\theta|$; $u = (u_1, u_2, u_3)$ where $u_1(x) = |x_1|$, $u_2(x) = |x_2|$, and $u_3(x) = 1$ if $x_1x_2 > 0$ and $= 0$ otherwise; and $u_s = |s|$ or $v(x) = u_s s(x) = |x_1 + x_2|$. We prove it for u only: (i) $u(g^\pm x) = u(x)$; (ii) $u(x) = u(x')$ implies $x_i = \pm x_i'$ with the same sign for $i = 1, 2$, that is, $x = g^+x'$ or $x = g^-x'$. Thus the magnitude of each coordinate x_i, together with the knowledge of which pair of diagonally opposite quadrants contains (x_1, x_2), form a maximal invariant u (a diagram may be helpful).

The theorem states that any conditional probability statement about any invariant function t—that is, any function for which $t(x) = t(-x)$—given v is free of θ-dependence; i.e., among invariant functions, $|x_1 + x_2|$ contains all the available information about $|\theta|$. For example, any statement about $|x_1|$, or about $(|x_1|, |x_2|)$, given $|x_1 + x_2|$ is free of dependence on $|\theta|$. This example is readily extended to samples of size $n > 2$.

For a second example, suppose $x = (x_1, \cdots, x_n)$ where the x_i's are independent and normal with $\theta = (\mu, \sigma)$, Θ is the upper half-plane, and G is the group of scale changes, an element of which multiplies each x_i by a specific positive constant c (see pp. 98–99 in [13]). The sample mean and standard deviation together constitute a sufficient statistic, and the basic assumptions are readily verified (the induced groups $\bar{G}$ and G_s again being groups of scale changes). Assumption A is satisfied since the absolutely continuous measure with derivative $1/c$ is an invariant measure on the Borel sets of the positive reals $\{c\}$, and correspondingly on G. Alternatively, Assumption C may be verified, the set A_0 being the line on which $x_1 = x_2 = \cdots = x_n$ (see II, Example 7.1 and 7.3).

The maximal invariants are found to be $\gamma(\theta) = \mu/\sigma$; $u(x) = (x_1/x_n, \cdots, x_{n-1}/x_n, \operatorname{sgn} x_n)$; and v is Student's t-statistic. The theorem thus states that the t-statistic is sufficient for the class of distributions of u, this class being indexed by γ. If t' is any invariant statistic—e.g., the sample mean divided by the sample range—then the t-statistic is sufficient for the distributions of t' in the sense that any probability statement about $t'(x)$ given $v(x)$ is parameter-free.

Similarly, the magnitude (or square) of the t-statistic is invariantly sufficient for the distributions of the maximal invariant $u(x) = (x_1/x_n, \cdots, x_{n-1}/x_n)$ under changes in sign and scale, the distributions being indexed by γ^2.

Uniform and exponential location-scale parameter examples may be treated analogously.

I.4. Invariant sufficiency and transitivity. In this section we discuss sufficiency and invariance for a sequential experiment. The experiment may be

terminated at any stage, but performance of stage n implies previous performance of stages $1, 2, \cdots, n-1$.

We must distinguish three kinds of probability models:

(i) the *component* or *marginal models* $X_{n\theta} = (\mathfrak{X}_n, \mathfrak{A}_n, P_{n\theta})$ for the stage n data x_n $(n = 1, 2, \cdots)$,

(ii) the *joint (n-fold) models* $X_{(n)\theta} = (\mathfrak{X}_{(n)}, \mathfrak{A}_{(n)}, P_{(n)\theta})$ for the accumulated data $x_{(n)} = (x_1, \cdots, x_n)$ through stage n, and

(iii) the *sequential model* $X_\theta = (\mathfrak{X}, \mathfrak{A}, P_\theta)$ for the whole sequence of data $x = (x_1, x_2, \cdots)$.

Here, $\mathfrak{X}_{(n)}$ and $\mathfrak{X}$ are the product (n-fold and infinite, respectively) sample spaces with components $\mathfrak{X}_1, \mathfrak{X}_2, \cdots, \mathfrak{X}_n, \cdots$, and $\mathfrak{A}_{(n)}$ and $\mathfrak{A}$ are the respective product σ-fields of events [32]. For each $\theta \in \Theta$, P_θ is a probability measure on $(\mathfrak{X}, \mathfrak{A})$ and $P_{(n)\theta}$ and $P_{n\theta}$ are the corresponding joint and marginal probability measures derived therefrom; thus, for $A \in \mathfrak{A}_{(n)}$, $P_{(n)\theta}(A)$ is the probability according to P_θ that the n-tuple $x_{(n)}$, obtained by truncating the sequence x, lies in A, and similarly for $P_{n\theta}$.

We shall largely be concerned with the sequence of joint models $\{X_{(n)\theta}\}$. The concepts of sufficiency and transitivity are defined (below) in terms of this sequence. Invariance, however, is more suitably defined in terms of the sequential model X_θ, although, by giving up justification for invariance reductions, we could avoid the sequential model altogether.

If, for each n, s_n is sufficient for the class $X_{(n)\Theta}$ of joint models $(\theta \in \Theta)$, then $s = (s_1, s_2, \cdots)$ is called a *sufficient sequence* for X_Θ; s_n is a function of the first n observations.

For each n, suppose t_n is a function of $x_{(n)}$. If, for all θ and each n, the conditional distribution of t_{n+1} given $x_{(n)}$ is identical with the conditional distribution of t_{n+1} given t_n, then $t = (t_1, t_2, \cdots)$ is said to be a *transitive sequence* for X_Θ. This definition is adequate for the discrete and continuous case examples treated here; a general definition appears in II.4 and [4]. The idea is that all the information about t_{n+1} contained in $x_{(n)}$ is carried by $t_n = t_n(x_{(n)})$.

In sequential inference problems about θ, Bahadur [4] has shown that attention may be confined to sequential decision rules, here called *s-rules*, which depend at each stage n only on s_n, provided $s = (s_1, s_2, \cdots)$ is a sufficient *and* transitive sequence for X_Θ.

We now introduce an invariance structure on X_Θ. Suppose G is a group of transformations g on the sequential sample space $\mathfrak{X}$ for which $gX_\Theta = X_\Theta$ with maximal invariant γ on Θ. We shall further assume that each g induces a transformation $g_{(n)}$ on the n-fold sample space $\mathfrak{X}_{(n)}$, that is, if x_n' is the nth component of $x' = gx$ then $x'_{(n)} = g_{(n)}x_{(n)}$. It is easily seen that $g_{(n)}X_{(n)\Theta} = X_{(n)\Theta}$, that is, the joint models are also invariant (but see the next paragraph). In particular, this further assumption holds if g acts component-wise: $gx = (g_1x_1, g_2x_2, \cdots)$, and this commonly occurs in applications (see also [26]). Typically, the stages in the sequential experiment are mutually independent copies in which case g must act component-wise with identical components g_n—for example, each

$\mathfrak{X}_n$ is a Euclidean plane and each g_n is a rotation, $g_{(n)}$ rotating each of the n component planes the same amount.

(Strictly speaking, all identical joint (and marginal) models—which need not be distinct for all $\theta \in \Theta$ even though the sequential models are—should be given the same index value, by introducing a reduced parameter index $\theta_n = \omega_n(\theta) \in \Theta_n$ say; the maximal invariant on Θ_n may also vary with n, being a function of γ however. For example, suppose the x_n's are mutually independent $N(\mu_n, \mu_n^2)$, $\theta = (\mu_1, \mu_2, \cdots)$, and $\theta_n = (\mu_1, \cdots, \mu_n)$, a new parameter being introduced at each stage; let g act component-wise with $g_n x_n = c_n x_n (c_n > 0)$, and the maximal invariants under $\bar{G}$ and $\bar{G}_{(n)}$ are $\gamma = \{\operatorname{sgn} \mu_n\}$ and the first n components thereof, respectively. We choose to avoid this notational complexity, feeling the imprecision should cause no real difficulty.)

Now let u_n denote a maximal invariant on $\mathfrak{X}_{(n)}$ under $G_{(n)}$. Since, as is clear, $G_{(n)}$ induces $G_{(m)}$ for $m < n$, u_m considered as a function of $x_{(n)}$ (depending on the first m coordinates) is invariant and hence a function of u_n; that is, from knowledge of the value of one term in the sequence $u = (u_1, u_2, \cdots)$, all prior terms may be evaluated. However, u itself is not necessarily a maximal invariant under G; but it is this sequence of maximal invariants (under $G_{(n)}$) that is relevant in the sequential decision problem since a maximal invariant under G would depend on the whole sequence $x = (x_1, x_2, \cdots)$ which is not available to the decision-maker.

We therefore interpret the *principle of invariance in the sequential case* as stipulating that attention be confined to u-rules—that is, to decision procedures that depend at stage n only on the value of u_n. (This is consistent with the definition in [26] where g acts component-wise. Application of the invariance principle in the sequential decision problem presumes a cost function which is invariant under both G and $\bar{G}$, for example, constant cost per observation.) In effect, we replace the original sequence of joint probability models $\{X_{(n)\theta}\}$ with the sequence $\{U_{n\gamma}\}$ where $U_{n\gamma}$ is the model for u_n. A component-wise sufficiency reduction on u leads to a sequence $v = (v_1, v_2, \cdots)$ which may be called an *invariantly sufficient sequence* for X_Θ under G, each v_n being invariantly sufficient for $X_{(n)\Theta}$ under $G_{(n)}$. Hence, when invoking invariance, restriction to v-rules is justified so long as v is transitive for the sequence of models $U_{n\gamma}$.

The Stein Theorem provides an alternative means of reduction from the sequence x to the sequence v, assuming $G_{(n)}$ induces a group of transformations on the sample space of s_n. The theorem asserts (under certain assumptions) that a maximal invariance reduction applied component-wise to s leads to an invariantly sufficient sequence v. (The diagram of I.1 is relevant only if we append subscripts (n) to X and U and subscripts n to S and V.)

One problem then remains—that of verifying the transitivity of v. Fortunately, as shown in II.4, it is sufficient to verify the transitivity of s, so that the upper route is completely justified; that is, we may make a sufficiency reduction from $x_{(n)}$ to s_n, verify the transitivity of s, and then make a maximal invariance reduction from s_n to v_n, and we will obtain an invariantly sufficient and transi-

tive sequence v. The reason this is permissible is that, in proving the Stein Theorem in Part II, a stronger result is actually obtained; this result may be roughly described as asserting the conditional independence of s_n and u_n given v_n—i.e., to the factoring of the conditional joint distribution of s_n and u_n into the conditional distribution of s_n and the conditional distribution of u_n—and this result can be used to show that transitivity of s implies transitivity of v.

Finally then, a problem of interest in applying the Stein Theorem to obtain v-rules is the verification of the transitivity of the sufficient sequence s. The following result (see Theorem 4.3 in II) is quite useful for this purpose. Suppose the original random variables with values $x = (x_1, x_2, \cdots)$ are mutually independent; then s is a transitive sequence if s_{n+1}, a function of $x_{(n+1)} = (x_{(n)}, x_{n+1})$, is a function of $s_n(x_{(n)})$ and x_{n+1} only, i.e., if s_{n+1} depends on $x_{(n)}$ only through s_n. This condition is easily verified for exponential class laws, where s_{n+1} is of the form $s_n + f(x_{n+1})$, and for nonparametric problems where sets of ordered observations constitute the sufficient statistic at any stage. In particular, this condition holds in all examples discussed in this paper (I.5 8).

I.5. Application to sequential tests of composite hypotheses: v-rules. Cox [9] has proposed that sequential tests of simple hypotheses about a parametric function γ, which are composite hypotheses about θ, may be obtained by applying a SPRT—or any *generalized sequential probability ratio test* (GSPRT) [27] for that matter—to a sequence of statistics whose distributions depend only on γ. (It should be noted that Wald [45] and Barnard [5] have proposed alternative approaches to deriving sequential tests of such hypotheses; see also [25], [14], [36].) In the framework of the previous section, an invariantly sufficient and transitive sequence v is such a sequence, and restriction to v-rules is simply a consequence of invoking the invariance principle; restriction to SPRT's applied to $v_1, v_2, \cdots$, which turn out to be v-rules, is on the other hand largely a matter of convenience.

Since v_n is sufficient for the distributions of any invariant function of which v_n is a function (see remarks after definition of invariant sufficiency in I.2), v_n is sufficient for the distributions of $v_{(n)} = (v_1, \cdots, v_n)$. Therefore, the joint density (with respect to a suitable dominating measure) of $v_{(n)}$ factors according to the Fisher-Neyman factorization theorem for sufficient statistics. The ratio of densities of $v_{(n)}$ at fixed values of γ, say γ_1 and γ_0, on which any GSPRT is based, thus reduces to the ratio of densities of v_n at γ_1 and γ_0; hence a GSPRT based on v depends only on v_n, not $v_{(n)}$, at stage n, and is thus a v-rule. The (joint) density of $v_{(n)}$ need not be known since only the (marginal) density of v_n is required. This factorization is the essence of Cox's theorem.

Actually, Cox's theorem is incompletely stated ([9], [22], [14]). The invariance assumption $gX_\Theta = X_\Theta$ is not explicitly assumed by Cox, but used in the proof (to establish the last line on p. 291 of his paper). That the theorem is invalid without this assumption is demonstrated by the following counterexample (Cox's notation): (x_1, x_2, x_3) are independently normal with unit variances and means $(\theta_1, \theta_1 + \theta_2, 0)$, $t_1 = x_1$, $t_2 = x_2$, $u = x_2 + x_3$, and the transforma-

tions take (x_1, x_2, x_3) into $(x_1, x_2 + c, x_3 - c)$. Then t_1 and u are independent so that their joint density factors, but the second factor, the $N(\theta_1 + \theta_2, 2)$ density of u, involves θ_1. In this example the model is not invariant under the transformations, and the conclusion of Cox's theorem is invalid although his conditions (as we understand them) are satisfied.

Our basic assumptions and Assumption C provide a corrected version of Cox's conditions; however, it is usually simpler to verify Assumption A than Assumption C, and, through Assumption B, an extension to some nonparametric applications is also made possible. Moreover, the justification for confining attention to v-rules is made precise and this enables further investigation of the properties of sequential procedures based thereon. Thus our approach corrects, simplifies, extends and motivates the application of Cox's technique.

Since the v_n's are not independent and identically distributed, SPRT's applied to them do not, in general, have any known optimum property (but see the end of I.7).

We list under four headings below the properties that are known:

(i) *strength:* For tests of simple hypotheses about γ, use of Wald's boundaries $B = \beta/(1 - \alpha)$ and $A = (1 - \beta)/\alpha$ provides approximate upper bounds (α, β) on the true error probabilities, whatever the values of the nuisance parameters. [It is frequently suggested (e.g., [9], [10] and pp. 98 and 250 in [30]) that in order to use Wald's boundaries for a SPRT one must prove termination with certainty. However, it is easily verified that the requirements on the error probabilities are fulfilled as approximate upper bounds, rather than approximate equalities, whether or not termination is certain. (The word "approximate" before "upper bounds" is really only justified if the error probabilities are small, but may in fact be deleted throughout if Wald's conservative boundaries $B = \beta$ and $A = 1/\alpha$ are used.)]

(ii) *termination:* Many such procedures, including those in which v_n has a *monotone likelihood ratio* (MLR), may be shown to terminate with certainty (for all θ or even more generally). For some specific examples, termination with probability one has been known for some time (e.g. [10]); more recently, rather general termination results were obtained by Wirjosudirdjo [17], Ifram [21] and Berk [6]. One or more of these references provides proof of termination for all examples in this paper (though only under H_0 and H_1 in the rank test example). When termination under H_0 and H_1 is assured, the approximate bounds in (i) become approximate equalities.

(iii) *OC-function:* The *operating characteristic functions* of these tests—or of any GSPRT's of γ_0 vs. γ_1 applied to v—depend on θ only through γ, and, if v_n has a MLR for γ in Γ, the OC-functions are monotone in γ (real) [15]; this occurs in most normal theory and exponential class examples. Thus, assuming $\gamma_0 < \gamma_1$, these tests are effectively tests of the hypotheses $\gamma \leqq \gamma_0$ vs. $\gamma \geqq \gamma_1$. But approximations to the OC-functions are not generally available. (Wald [45] has given a monotonicity theorem but his proof is incomplete. In the problem considered by him, he claims that it is sufficient to prove that the density

ratio at γ_1 and γ_0 is monotone in v_n but he did not verify this claim. Also, the proof of a similar monotonicity theorem in [29] is invalid for tests based on a sequence of dependent variables, such as v. A valid monotonicity theorem for the case of independent variables, such as the z-rules of I.7, appears in [29] and [30]. Valid theorems for the dependent case appear in [15] and [16] and, implicitly, in [47].)

(iv) ASN-*function:* Little is known about the *average sample number functions* for these tests except for some heuristic approximations supported by empirical investigations (see [25], [35], [2], [22]). The results of Ifram [21] may be used to obtain alternative approximations.

We now consider the examples introduced in I.3. First, suppose the observations are independent $N(\theta, 1)$ and we wish to test $|\theta| \leqq \gamma_0$ against the two-sided alternative $\theta \geqq \gamma_1$ $(> \gamma_0 \geqq 0)$. The class of probability models is invariant under the group of sign changes for every n, as are the hypotheses. In I.3 we found $v_n = |x_1 + \cdots + x_n|$ to be invariantly sufficient for the joint models of the first n observations, and $v = (v_1, v_2, \cdots)$ is then found to be an invariantly sufficient and transitive sequence; restriction to v-rules is then justified under the principle of invariance.

A SPRT of γ_0 vs. γ_1 based on v depends at stage n only on the ratio of densities of v_n (using the sufficiency factorization), and this is readily found to be

$$\exp\left[-n(\gamma_1^2 - \gamma_0^2)/2\right] \cosh(v_n\gamma_1)/\cosh(v_n\gamma_0).$$

Sampling is continued as long as the ratio remains between Wald's boundaries B and A (or B_n and A_n for a GSPRT). Since v_n may be verified to have a MLR in $\gamma = |\theta|$, the OC-function —which depends only on γ— is monotone in γ, and the test terminates with certainty for all γ [21] (and even more generally [6]). The true error probabilities are approximately equal to the prescribed ones.

This example has application in the sequential testing for the significance of the difference between two means against two-sided alternatives when the observations are normal with equal and known variances, a difference being observed at each stage of sampling. The Sobel-Wald [43], Armitage [3], and Schneiderman-Armitage [40] sequential test procedures are alternative to the one above, and not based on invariance theory, but the symmetric version of each is still a v-rule, and in fact a GSPRT based on v.

Consideration of the second example of I.3 would lead to the WAGR one-sided sequential t-test of $\mu/\sigma \leqq \gamma_0$ vs. $\mu/\sigma \geqq \gamma_1$ with v_n equal to the t-statistic (or a monotone function thereof) based on the first n observations ([10], [36] and p. 250 of [30]); its properties are analogous to those above (i)–(iv). We omit further consideration of it, but consider the two-sided case—the variance unknown analog of the first example above—in the next section; see also I.7. That these t-tests have a broader applicability may be seen from Example 7.3 in II.7.

Sequential tests about one or two normal variances and sequential tests about normal correlation coefficients may be derived analogously.

I.6. Sequential F, t^2, and T^2 tests. Sequential F-tests and T^2-tests appear to have been introduced by Stein [44]. Sequential F-tests were also considered by Nandi [33]. Johnson [23] invoked Cox's method and Hoel [20] employed the weight function method of Wald [45] to justify sequential F-tests; Hoel also pointed out invariance properties of them. F-tests are also treated in [35], [38], and [14]. Cox's method has been recently applied to sequential T^2-tests (and χ^2-tests) by Jackson and Bradley [22]. The two-sided t-test, or t^2-test, is treated here as a special case of the F-test (or the T^2-test); it appears in [34] and [37]. Some comparisons with alternative t^2-tests appear in [41]. Sequential tests for components of variance problems ([24], [14]) may also be derived through sufficiency and invariance considerations but will not be considered here. Our purpose here is to show in outline how these tests (or slight extensions thereof) can be derived from sufficiency and invariance considerations; we thus provide a rigorous basis for justifying and interpreting them. Alternative tests may be constructed by the methods of I.7.

F-test for fixed-effects model: We consider a sequential experiment in which each stage consists of a replication of a fixed-effects linear model experiment. We assume that the data from each stage are separately reduced to canonical form as described in Section 7.1 of [30] so that stage n yields data $x_n = (x_{n1}, \cdots, x_{np})$, where the x_{ni}'s are independent normal observations with common (unknown) variance σ^2 and with means $\theta_1, \cdots, \theta_l, 0, \cdots, 0$, $(l \leqq p)$. The hypotheses to be tested are $H_0 : \gamma \leqq \gamma_0$ vs. $H_1 : \gamma \geqq \gamma_1 (\gamma_1 > \gamma_0 \geqq 0)$ where $\gamma = \sum_{i=1}^{k} \theta_i^2/\sigma^2$, $k \leqq l$. If γ_0 is taken to be zero the null hypothesis may be described as $\theta_1 = \cdots = \theta_k = 0$; however, since such a hypothesis is usually known to be false *a priori* it may be more reasonable to assign a type I error bound to a larger parameter set, and a γ-interval is a mathematically convenient choice.

The sequential F-test for this problem will now be briefly described and justified. Many of the arguments sketched below can be verified in analogy with results in [30].

We define a group G of transformations g which act component-wise in an identical way on the canonical forms of the respective stages of the experiment. Each transformation is defined by an arbitrary positive number b, an arbitrary orthogonal matrix C, and $l - k$ arbitrary numbers $a_{k+1}, \cdots, a_l$, and transforms the stage n data as follows:

$$(x_{n1}, \cdots, x_{nk}) \to b \cdot (x_{n1}, \cdots, x_{nk}) \cdot C,$$

$$(x_{n,k+1}, \cdots, x_{nl}) \to b \cdot (x_{n,k+1} \dot{+} a_{k+1}, \cdots, x_{nl} \dot{+} a_l),$$

$$(x_{n,l+1}, \cdots, x_{np}) \to b \cdot (x_{n,l+1}, \cdots, x_{np}).$$

This group leaves the model for the data through stage n invariant, with γ (defined above) as a maximal invariant on the parameter space.

The sample means $\bar{x}_{n1}, \cdots, \bar{x}_{nl}$ of the first l components of the observations through stage n together with the conventional error mean square E_n (based on $\nu_n = (n-1)l + n(p-l)$ degrees of freedom) constitute a stage n sufficient

statistic, say s_n, and $s = (s_1, s_2, \cdots)$ is a sufficient and transitive sequence. The following transformation is induced on the sample space of s_n:

$$(\bar{x}_{n1}, \cdots, \bar{x}_{nk}) \to b \cdot (\bar{x}_{n1}, \cdots, \bar{x}_{nk}) \cdot C$$

$$(\bar{x}_{n,k-1}, \cdots, \bar{x}_{nl}) \to b \cdot (\bar{x}_{n,k-1} + a_{k+1}, \cdots, \bar{x}_{nl} + a_l)$$

$$E_n \to b^2 \cdot E_n .$$

A maximal invariant under this induced group is the F-statistic F_n (with k and ν_n degrees of freedom) conventionally used to test the hypothesis $\gamma = 0$ (or $\gamma \leqq \gamma_0$) based on *all* the data available through stage n; thus F_n is the ratio of the mean squares due to hypothesis and error based on the first n replications.

Since every almost invariant function of s_n is known to be equivalent to an invariant function, Assumption A holds, and the Stein Theorem leads to the conclusion that $F = (F_1, F_2, \cdots)$ is an invariantly sufficient (and transitive) sequence; hence, F_n is sufficient for the distributions of any invariant function of which it is a function, specifically for the distributions of $(F_1, \cdots, F_n)$. (Application of the Stein Theorem could also be validated by verifying Assumption C, in analogy with Cox's theorem, but the verification is both tedious and unnecessary.) The likelihood ratio for $(F_1, \cdots, F_n)$ at γ_1 and γ_0 then reduces to the likelihood ratio for F_n at γ_1 and γ_0, that is, the ratio of two non-central F-densities. This ratio may be conveniently expressed as $R_n = h_n(\gamma_1)/h_n(\gamma_0)$ where

$$h_n(\gamma) = \exp(-n\gamma/2) M[(\nu_n + k)/2, k/2; (n\gamma/2) \cdot z_n/(1 + z_n)],$$

$M(\cdot, \cdot; \cdot)$ is the confluent hypergeometric function and $z_n = F_n k/\nu_n$, the ratio of sums of squares due to hypothesis and error [38]. A SPRT based on the ratios $\{R_n\}$ has properties (i)–(iv) listed in 1.5 since the non-central F-statistic has a M.L.R.

References to available tables for the confluent hypergeometric function may be found in [42]; asymptotic expansions given there may also be used to develop approximate procedures. See also [35], [36], [38], [39] in these regards.

Note that we do not reduce the data available through stage n to canonical form, but only the data of each stage separately. This is essential in our formulation to permit a consistent component-wise group structure. Actually the successive stages need not be perfect replications of one basic experiment so long as the canonical forms are perfect replications; in fact, we can permit any number of the last $p - k$ components, or all of the first k components, of the data in canonical form to be missing at any stage by making only minor alterations above. Indeed, p and the number $(l - k)$ of nuisance parameters may be infinite, provided only that each row in the design matrix have a finite number of non-zero entries. Thus, new nuisance parameters may be introduced at each stage to adjust for suspected stage-to-stage effects. (We call the stages replications, but they include any time effects.) Some accounting for such effects, even stage-wise variation in σ^2, is also possible by using an alternative procedure, namely an SPRT based on a sequence of independent F-statistics, one computed from each stage (assuming

$l < p$). Such procedures are considered in I.7 (specifically in the context of t-tests rather than F-tests).

Finally, the sequential F-test may be shown to be valid in any situation in which the non-sequential F_n-test, based on the accumulated data through stage n, has a power function depending on γ only. In this case, one can transform to new observations x_n', where $x'_{(n)}$ is a function of $x_{(n)}$, and in terms of the new observations the first k components are replicated equally, or not at all, at each stage so that the preceding analysis is applicable for the primed data.

Two-sided t-test: The above F-test reduces to a two-sided t-test, or t^2-test, when $k = 1$; we further assume below that $p = l = 1$. (The case $k = 1$, $p = l = 2$ is discussed in [17].) Thus, based on a single sequence of observations $x_1, x_2, \cdots$, from a normal population with unknown mean μ and unknown variance σ^2, we wish to test $H_0 : (\mu/\sigma)^2 \leqq \delta_0^2 (= \gamma_0)$ against $H_1 : (\mu/\sigma)^2 \geqq \delta_1^2 (= \gamma_1)$. (The transformations take each x_i into $\pm bx_i$ and thus constitute changes in sign and scale; see I.3.) The likelihood ratio after n observations is as given above for the F-test with $\nu_n = n - 1$, $k = 1$, and F_n the square of Student's t-statistic based on $n - 1$ degrees of freedom; more conveniently, $z_n/(1 + z_n) = (\sum_{i=1}^n x_i)^2 / \sum_{i=1}^n x_i^2 = r_n$, say. When $\delta_0 = 0$, the SPRT reduces to the common WAGR two-sided t-test treated in [34] and [37]. The *restricted* [3] and *wedge* [41] procedures are alternative GSPRT's based on the same invariantly sufficient sequence of t^2-statistics. See also I.7. (*Note:* Using Wald's weight function one can obtain a similar test with the modification of reducing by unity the first argument of the confluent hypergeometric function. However, even for $\delta_0^2 = 0$, the case considered by Wald, we know of no rigorous proof of Wald's inequalities on the two error probabilities; for the kind of arguments required, see [5].)

Tables [34] are available for carrying out the SPRT when $\delta_0 = 0$. Otherwise, tables of the confluent hypergeometric function are required (see above). Alternatively, one can approximate the confluent hypergeometric function to obtain a simpler form of the test. Following Rushton [36], one obtains the following approximate sampling procedure, using his simplest approximation: continue sampling only if $a_n < r_n < b_n$ where $a_n = n\{[1 + 2\lambda^2 + (\lambda/\rho)(a/n)]^2 - 1\}^2/\lambda^2$ and $a = \log A$, $\lambda = (\delta_1 + \delta_0)/2$, $\rho = (\delta_1 - \delta_0)/2$, and similarly for b_n with a replaced by $b = \log B$. Better approximations or the exact formulas could be used whenever r_n lies close to a critical value a_n or b_n.

T^2-test: The same approach as for the F-test above could be carried through for multivariate linear models, as in Section 7.9 of [30], but a single (numerical) maximal invariant on the space of the sufficient statistic or on the parameter space is usually not available; instead, the roots of certain determinantal equations play these roles. Sequential tests of simple hypotheses about these parametric roots could be carried out, but such hypotheses would seem to be of little practical interest; there is no available sequential analog of the maximum root, trace, or likelihood ratio tests for this problem (but see I.7). However, whenever, there is a single non-zero root, the problem reduces to that of a sequential T^2-test, an important special case of which we outline below. (See [30] and [1].)

Consider two p variate multinormal populations with equal (and unknown) non-singular dispersion matrices Σ and unknown mean vectors θ_1 and θ_2. At each stage of the experiment an observation from each population is drawn independently, say x_{n1} and x_{n2} (p-vectors). We wish to test the hypotheses $\gamma \leqq \gamma_0$ vs. $\gamma \geqq \gamma_1 (\gamma_1 > \gamma_0 \geqq 0)$ where $\gamma = (\theta_1 - \theta_2)'\Sigma^{-1}(\theta_1 - \theta_2)$. With $\gamma_0 = 0$, this is the problem of sequentially testing equality of two mean vectors.

Let g be a component-wise transformation with identical components taking x_{ni} into $(x_{ni} + a)\cdot L (i = 1, 2)$ where a is an arbitrary vector of constants and L is an arbitrary non-singular matrix. These transformations leave the problem invariant, and γ is a maximal invariant on the parameter space. The two vectors of sample means, $\bar{x}_{n1}$ and $\bar{x}_{n2}$, and the pooled sample dispersion matrix D_n form a sufficient statistic based on the data through stage n; the sequence of sufficient statistics is transitive, and a transformation is induced on the sufficient statistic taking $\bar{x}_{ni}$ into $(\bar{x}_{ni} + a)L (i = 1, 2)$ and D_n into $L'D_nL$; moreover, Assumption A holds. A maximal invariant is T_n^2, the non zero root of the determinantal equation $|A_nD_n^{-1} - \lambda I| = 0$ where $A_n = (\bar{x}_{n1} - \bar{x}_{n2})'(\bar{x}_{n1} - \bar{x}_{n2})$. ($T_n^2$ is an arbitrary constant for $n \leqq (p + 1)/2$.) Hence, $T^2 = (T_1^2, T_2^2, \cdots)$ is an invariantly sufficient and transitive sequence and may be used to construct sequential tests about the parameter function γ.

Since T_n^2 is a Hotelling T^2-statistic (with $2(n - 1)$ degrees of freedom), its distribution is essentially that of non-central F; hence, the stage n likelihood ratio is of the same form as in the F-test, with $k = p$, $\nu_n = 2(n - 1) - p + 1$, and $z_n = T_n^2/2(n - 1)$. When $p = 1$, this reduces to the two-sample t^2-test [17].

Similarly, an invariantly sufficient and transitive sequence of Hotelling T^2-statistics (with $n - 1$ degrees of freedom) may be constructed for testing hypotheses about $\gamma = \theta_1'\Sigma^{-1}\theta_1$ when sampling from a single p-variate normal population with mean vector θ_1 and dispersion matrix Σ; the transformations then are of the form $x_{n1} \to x_{n1}\cdot L$.

If Σ is assumed known in these problems, one may obtain analogously *sequential χ^2-tests* [22].

I.7. Alternative sequential tests: z-rules. An alternative approach to the construction of sequential tests about a parameter γ, when the successive stages of the sequential experiment are mutually independent, is as follows: let t_n be a statistic based on the stage n data (a function on $\mathfrak{X}_n$) whose distribution depends only on γ, and let z_n be a function of $t_{(n)} = (t_1, \cdots, t_n)$ which is sufficient for the distributions of $t_{(n)}$. Then a GSPRT of simple hypotheses about γ can be based on the t sequence, or equivalently (by sufficiency) on the z-sequence. We call such tests *z-rules*, decisions at stage n depending only on the value of z_n. An advantage of z-rules is that the t_n's are mutually independent and possibly identically distributed as well, in which case Wald's ASN and OC approximations and his termination proofs are applicable; Lehmann's [30] monotonicity theorem for the OC function may also be applicable. However, unless such procedures are also v-rules (see below), they waste pertinent information about γ and so are presumably less efficient that z-rules; in fact, they may perform no better,

as far as the ASN is concerned, than nonsequential tests of equal strength (see [24]).

To derive z-rules from sufficiency and invariance considerations, we first note that, in the framework of I.4, G induces a group G_n on $\mathfrak{X}_n$ with elements g_n. Let w_n be invariantly sufficient for the component models $X_{n\theta}$ under G_n. For z-rules, invariance reductions (to w_n) are made separately stage-wise, and then sufficiency reductions yield the z-sequence; for v-rules, invariance reductions (to u_n) are made on the accumulated data, and then sufficiency reductions yield the v-sequence.

Now it may be necessary to group the successive stages of the experiment in order that the w-sequence does not degenerate into a sequence of constants. As an example, consider a one-sided t-test situation (see I.5 and I.3) in which observations are taken in successive groups of size k (>1) from a normal population. Student's t-statistic calculated from the stage n data ($k-1$ degrees of freedom) is invariantly sufficient under scale changes on the stage n data and $\gamma = \mu/\sigma$. Thus w is a sequence of independent t-statistics. An SPRT of γ_0 vs. γ_1 based on w is easily constructed; this is a z-rule (we can let $z_n = w_{(n)}$). After a total of nk observations this z-rule would utilize n t-statistics with a total of $n(k-1)$ degrees of freedom; in contrast, the WAGR t-test, which is a v-rule, would utilize one t-statistic with $nk-1$ degrees of freedom. Hence, the z-rule wastes $n-1$ degrees of freedom (the between stages degrees of freedom); also, it only permits termination after multiples of k observations. On the other hand, approximations to its OC and ASN functions are available.

A compromise between these two approaches would retain some of the advantages of each; specifically, observations may be taken one at a time and, after $nk+m$ observations ($1 < m \leqq k$, $n = 0, 1, \cdots$), decisions based on the probability ratio of the mutually independent t-statistics $t_1, t_2, \cdots, t_n, t'_{n+1}$, each of the first n t_m's being based on $k-1$ degrees of freedom and t'_{n+1} being based on $m-1$ degrees of freedom (termination is not permitted when $m = 1$).

Examples of z-rules in the literature include a range test for normal variances introduced by Cox [8], some sequential tests for variance components introduced by Johnson [24], and two sequential rank tests for the two-sample problem proposed by Wilcoxon, Rhodes and Bradley [46] (see I.8). Other possibilities are abundant; for example, a group sequential test for the multivariate linear hypothesis could be constructed based on a sequence of independent maximum root statistics, trace statistics, or likelihood ratio statistics.

Now it may happen, though not in any of the examples considered so far in this paper, that a z-rule is in fact a v-rule. This occurs whenever $w_{(n)} = (w_1, \cdots, w_n)$ is a maximal invariant under $G_{(n)}$, that is, $w_{(n)} = u_n$. This is so in particular if $G_{(n)}$ is isomorphic to $G_1 \times G_2 \times \cdots \times G_n$. An example of this is provided by a modification of the t-test example above. Suppose the k observations in stage n are independent $N(\gamma\sigma_n, \sigma_n^2)$; thus the mean and standard deviations (all unknown) may now vary from stage to stage but the ratio remains constant. Let $g = (g_1, g_2, \cdots)$ where g_n applies a scale change to the stage n

data, and g_n may now vary with n. Then $w_{(n)}$ above is easily seen to be invariantly sufficient under $G_{(n)}$ and w_n under G_n. Thus, under these conditions, an SPRT based on the sequence of independent t-statistics is both a z-rule and a v-rule. The sequential F-test and other normal theory examples could be modified analogously.

When $w_{(n)}$ does coincide with u_n, SPRT's constructed from the w-sequence (as in the t-test example above) are easily seen to have a Wald-type optimality among all invariant procedures, that is, the ASN is minimized for all θ such that $\gamma(\theta) = \gamma_0$ or γ_1 among all invariant procedures with the same or smaller error probabilities. We conclude this section with a second example; a third example will be given in the next section. (If the w_n's are Fraser-sufficient as well as invariantly sufficient, as in the example below, then the restriction to invariant procedures in this optimality may be removed; see [16] and [12].)

Suppose stage n of the experiment yields two independent normal observations with unit variances and unknown means $\mu_n + \gamma$ and μ_n, respectively. Suppose G is the group of transformations which adds an arbitrary constant a_n to both observations from stage $n(=1, 2, \cdots)$. Then w_n may be taken as the difference between the stage n observations, and $u_n = (w_1, \cdots, w_n)$ is a maximal invariant under $G_{(n)}$ as is γ under the induced group on the parameter space. (The same statements hold if μ_n and a_n are constants μ and a.) Then $z_n = v_n = \sum_{i=1}^{n} w_i$. An SPRT of simple hypotheses about γ based on the w-sequence has the Wald optimal property.

I.8. Nonparametric sequential applications. In a sequential test (at least in a SPRT), the two hypotheses and two kinds of error are treated in a similar fashion. In most nonparametric tests, however, the alternative hypotheses are typically rather vague or all-encompassing. Thus, to obtain a sequential nonparametric test from available theory, one must consider rather specific alternatives—sufficiently specific so that the probability distribution of some test statistic (perhaps invariantly sufficient) is completely specified by the alternative hypothesis, as it is by the null hypothesis. The practicality of such a specification is perhaps rare.

One such example, however, is the *sign test* for which one reduces nonparametric measurement data to binomial data by classifying successive independent observations (or pairs of observations) simply as "successes" or "failures" (see pp. 147–149 in [30]). Sequential binomial tests of whether the success probability is large or small can then be performed. Sometimes such data reductions may be justified by invariance considerations, e.g., in *paired comparisons* (p. 220 in [30]) or when testing for *symmetry* (p. 242 in [30]); see also the example below. Here, the lower route is the convenient one—reducing by invariance and then by sufficiency—so that the Stein Theorem is not required. However, when testing two-sided hypotheses which are symmetric about $\frac{1}{2}$ in the success probability p, invariance may again be applied after sufficiency to reduce the data further to the magnitude of the deviation of the proportion of success from $\frac{1}{2}$; here $\gamma = |p - \frac{1}{2}|$. Then the two-sided sequential binomial test may be derived, using the methods

of I.5. Both of these sequential sign tests have properties (i)–(iv) described in I.5.

Some examples from the theory of most powerful rank order tests [28] can also be handled. In this theory (nonsequential), one does specify a particular type of alternative against which maximum power is desired, for a test of level α. A sequential test of specified strength (α, β) can, at least in principle, be based on a sequence of most powerful test statistics, calculated stage-wise or from the accumulated data. When invariance considerations are applicable, the Stein Theorem may facilitate such tests. We exemplify this with a two-sample sequential rank test; an analogous one-sample rank test of symmetry about the origin may also be developed.

Suppose at each stage of experimentation an observation is taken independently from each of two populations with (unspecified) continuous distribution functions F and F', respectively, and one wishes to test $H_0 : F' = F$ against $H_1 : F' = F^2$ (see [28] or [13]). This is a particular case of the "one variable is stochastically larger than the other" alternative. Letting $r_1, r_2, \cdots, r_n$ be the ordered ranks (increasing order) of the observations from the second population from a combined ranking of all the data available through stage n, we find that $v_n = (r_1, \cdots, r_n)$ is invariantly sufficient (and transitive) under the group G, an element of which applies an identical monotone continuous transformation to each observation (see remarks on Assumption B in I.2). Using the Stein Theorem and [28], the probability ratio at stage n is found to be

$$2^n r_1(r_2 + 1) \cdots (r_n + n - 1)/(2n + 1)(2n + 2) \cdots (3n),$$

and a SPRT is readily performed by comparing this ratio at each stage with Wald's boundaries A and B. Similar results are available when the alternative is $F' = F^k (k > 1)$ or $F' = h(F)$ for specified $h(\cdot)(F' \leqq F)$; also, sampling in pairs is not essential (see final paragraph). Nothing is known about the properties of these tests other than that specified bounds (approximate equalities) on the error probabilities are met (i), and termination occurs with certainty if either of the two hypotheses is true (ii); the latter follows from a theorem of Wirjosudirdjo [47], but whether termination is certain under other hypotheses is not known. However, under the invariance principle, any good test must be a v-rule, and these tests are v-rules. An alternative group sequential procedure, a z-rule, will be given in the last paragraph below.

Now let us consider a variation on the above example. We replace F and F' by F_n and F'_n in the assumptions and hypotheses, and no longer require the components g_n of $g = (g_1, g_2, \cdots)$ to be identical; that is, the pair of observations from stage n, having distributions F_n and F'_n, are both transformed by the same monotone continuous transformation, but the transformations and distributions may vary with n. The required invariance assumptions still hold. Letting $w_n = 1$ or 0 according as the sign of the difference between the stage n observations is positive or negative, we find that w_n is a maximal invariant under G_n and $u_n = (w_1, \cdots, w_n)$ is a maximal invariant under $G_{(n)}$. A sequential binomial test

or sign test (of $p = \frac{1}{2} = P(w_n = 1 \mid H_0)$ against $p = \frac{1}{3} = P(w_n = 1 \mid H_1)$) is then a v-rule *and* a z-rule and has the Wald optimal property among invariant procedures.

Finally, suppose k_n and k_n' independent observations are to be taken at stage n, if stage n is performed, from populations with distribution functions F_n and F'_n, respectively. (The sequences of numbers k_n and k_n', $n = 1, 2, \cdots$, are non-negative integers, arbitrarily determined but not dependent on the observations.) Thus, the distributions may vary from stage to stage but not within stage. The group G has a parallel structure: the component g_n of g transforms each observation from stage n by the same monotone continuous transformation but the g_n's may vary with n. Letting w_n denote the ordered ranks from the second (primed) population from a combined ranking of only the stage n data, a SPRT based on the w-sequence is both a v-rule and a z rule and has the Wald optimal property among invariant procedures. (At stage n the probability ratio is given by the product of the stage-wise probability ratios, and each of them is of the form $2^{k'} r_1(r_2 + 1) \cdots (r_{k'} + k' - 1)/(k + k' + 1)(k + k' + 2) \cdots (k + 2k')$.) However, if the hypotheses (and transformations) do not permit variation from stage to stage ($F_n = F$, $F'_n = F'$ and the components of g are identical), then this procedure is still a valid z-rule but not a v-rule; it terminates with certainty and has a known ASN (approximate), but presumably is less efficient than the v-rule which would re-rank all the available data at each stage rather than rank only within stages. This z-rule, and an analogous z-rule based on the rank sum $\sum r_i$ (there is no rank sum v-rule since the rank sum is not invariantly sufficient), have been proposed by Wilcoxon, Rhodes and Bradley [46], and designated the *configural rank test*.

PART II: GENERAL THEORY

II.0. Summary. $\mathcal{P}$ is a family of distributions on a σ-field $\mathfrak{A}$ of subsets of $\mathfrak{X}$, $\mathfrak{A}_S$ is a sufficient subfield of $\mathfrak{A}$, G is a group of invariance transformations g on $\mathfrak{X}$, $\mathfrak{A}_I$ is the σ-field of invariant members of $\mathfrak{A}$, and $\mathfrak{A}_{SI}$ is the intersection of $\mathfrak{A}_S$ and $\mathfrak{A}_I$. The main result establishes, under certain conditions, that $\mathfrak{A}_{SI}$ is sufficient for $\mathfrak{A}_I$. This is implied by the slightly stronger conclusion that $\mathfrak{A}_S$ and $\mathfrak{A}_I$ are conditionally independent given $\mathfrak{A}_{SI}$. Both conclusions have been established under any one of three assumptions. Assumption A is that $g\mathfrak{A}_S = \mathfrak{A}_S$ for each g, and that every $\mathfrak{A}_S$-measurable and almost invariant function is equivalent to an $\mathfrak{A}_{SI}$-measurable function. Assumption B is that there exists a conditional probability distribution Q such that $Q(gA, gx) = Q(A, x)$ for all $A \in \mathfrak{A}$, $x \in \mathfrak{X}$, $g \in G$. Assumption C is that $\mathcal{P}$ is a family of densities on n-space of the form $p_\theta(x) = g_\theta(s(x))h(x)$, the functions g_θ and h being positive, and that on some $\mathfrak{A}_{SI}$ set of probability 1 each transformation g is differentiable with Jacobian depending only on $s(x)$, $s(x) = s(x')$ implies $s(gx) = s(gx')$, s is differentiable with its matrix of partial derivatives of maximal rank, and $h(gx)/h(x)$ depends only on $s(x)$. A counter example shows that $\mathfrak{A}_{SI}$ need not be minimal sufficient for $\mathfrak{A}_I$ if $\mathfrak{A}_S$ is minimal sufficient for $\mathfrak{A}$. On the other hand, completeness of $\mathcal{P}$ on $\mathfrak{A}_S$ is

inherited by $\mathcal{A}_{SI}$. Some theorems concerning transitivity in the sequential case are given. Theorem 4.1 states that $\{\mathcal{B}_n\}$ being transitive for $\{\mathcal{A}_n\}$ is equivalent to $\mathcal{B}_n$ and $\mathcal{B}_{c(n,1)}$ being conditionally independent given $\mathcal{B}_{0n}$ for each n. Theorem 4.2 states that if, for each n, $\mathcal{A}_{3n} \subset \mathcal{A}_{1n} \subset \mathcal{A}_n$, $\mathcal{A}_{3n} \subset \mathcal{A}_{2n} \subset \mathcal{A}_n$, and $\mathcal{A}_{1n}$ and $\mathcal{A}_{2n}$ are conditionally independent given $\mathcal{A}_{3n}$, then $\{\mathcal{A}_{1n}\}$ transitive for $\{\mathcal{A}_n\}$ implies $\{\mathcal{A}_{3n}\}$ transitive for $\{\mathcal{A}_{2n}\}$. This implies (under Assumption A, B or C) that $\{\mathcal{A}_{SIn}\}$ is transitive for $\{\mathcal{A}_{In}\}$ if $\{\mathcal{A}_{Sn}\}$ is transitive for $\{\mathcal{A}_n\}$. Theorem 4.3 states that if $\mathcal{B}_1, \mathcal{B}_2, \cdots$ are independent subfields, $\mathcal{A}_n = \mathcal{B}_1 \vee \cdots \vee \mathcal{B}_n$ for each n, and $\{\mathcal{A}_{0n}\}$ any sequence such that $\mathcal{A}_{0n} \subset \mathcal{A}_n$ and $\mathcal{A}_{0(n+1)} \subset \mathcal{A}_{0n} \vee \mathcal{B}_{n+1}$, then $\{\mathcal{A}_{0n}\}$ is transitive for $\{\mathcal{A}_n\}$.

II.1. Introduction. The question with which this paper is concerned is phrased in Part I essentially as follows: If a sufficiency reduction and an invariance reduction of a problem are performed in succession, is the result independent of the order in which these two reductions are carried out? The purpose of Part II is to present a treatment of this question and its various solutions entirely in the language of subfields.

Let $\mathcal{X}$ be a space, $\mathcal{A}$ a σ-field of subsets of $\mathcal{X}$, $\mathcal{P}$ a family of distributions on $\mathcal{A}$, and $\mathcal{A}_S$ a sufficient subfield of $\mathcal{A}$. Let G be a group of invariance transformations g of $\mathcal{X}$ onto itself (precise definitions are given in Sections 2 and 3). Let $\mathcal{A}_I$ be the σ-field of invariant members of $\mathcal{A}$. The intersection $\mathcal{A}_S \cap \mathcal{A}_I$ will be denoted by $\mathcal{A}_{SI}$.

Suppose that $\mathcal{A}_S$ is induced by a sufficient statistic $(\mathcal{S}, \mathcal{A}', s)$, where s is a function from $\mathcal{X}$ onto $\mathcal{S}$ and $\mathcal{A}'$ is a σ-field of subsets of $\mathcal{S}$ such that $\mathcal{A}_S = s^{-1}(\mathcal{A}')$. The notion of a sufficiency reduction followed by an invariance reduction cannot be formulated very well unless every g induces a transformation in $\mathcal{S}$, i.e. $s(x_1) = s(x_2)$ implies $s(gx_1) = s(gx_2)$. In the subfield language this means that g transforms any member of $\mathcal{A}_S$ into a member of $\mathcal{A}_S$. We shall assume throughout that every $g \in G$ has this property (Assumption A (i)). It is clear that the inverse images under s of the invariant sets of $\mathcal{A}'$ constitute the subfield $\mathcal{A}_{SI}$.

An invariance reduction applied after a sufficiency reduction leads to a maximal invariant function on $\mathcal{S}$, say u (where u is supposed to be $\mathcal{A}'$-measurable). Since u induces the σ-field of invariant sets in $\mathcal{A}'$, the function $u(s(\cdot))$ on $\mathcal{X}$ induces $\mathcal{A}_{SI}$. The question stated in Part I is whether a maximal invariant function on $\mathcal{S}$ is *invariantly sufficient*, i.e. sufficient for $\mathcal{P}$ restricted to the invariant sets. Translated into the subfield language this question becomes: *Is $\mathcal{A}_{SI}$ sufficient for $\mathcal{A}_I$?*

It is not known whether the answer to the question in the preceding sentence is yes, in general, if only Assumption A (i) is made. C. M. Stein, in an unpublished manuscript (see Preface), was the first to recognize the problem, and to give sufficient conditions under which a maximal invariant function on $\mathcal{S}$ is invariantly sufficient. In the present paper we shall prove the desired result under various other sets of conditions, different from those of Stein. More specifically, we shall propose three different sets of conditions, called Assumptions A, B, and C. In applications it is convenient to have several possibilities to choose from, for, to give an example, conditions that are easy to check in situations involving normal

distributions may be hard to verify in nonparametric situations, and vice versa. It turns out that Assumptions A and C are usually easier to apply in parametric problems, B in nonparametric problems.

The sufficiency of $\mathcal{A}_{S|}$ for $\mathcal{A}_{\cdot}$ turns out to be a consequence of the following interesting relationship between the subfields $\mathcal{A}_S$, $\mathcal{A}_I$ and $\mathcal{A}_{\cdot|I}$: *$\mathcal{A}_S$ and $\mathcal{A}_I$ are conditionally independent given $\mathcal{A}_{\cdot|I}$*. This conditional independence obtains under any one of the Assumptions A, B or C. The formulation in terms of conditional independence of subfields has the advantage of being symmetric in $\mathcal{A}_S$ and $\mathcal{A}_I$, and in simplifying proofs, e.g. in Section 4.

Sequential aspects are treated in Section 4. At each sampling stage n we have the subfields $\mathcal{A}_n$, $\mathcal{A}_{Sn}$, $\mathcal{A}_{In}$, $\mathcal{A}_{S|n}$. Whether or not $\mathcal{A}_{S|n}$ is sufficient for $\mathcal{A}_{|n}$ depends only on the three subfields $\mathcal{A}_{Sn}$, $\mathcal{A}_{In}$ and $\mathcal{A}_{S|n}$, not on the whole sequence. However, an additional notion enters that does depend on the whole sequence, namely the notion of transitivity, introduced by Bahadur [4]. Questions of transitivity are answered in Section 4, relying heavily on the notion of conditional independence of subfields.

In Section 5, Assumption A is discussed. In the language of statistics Assumption A (ii) means that an almost invariant function on $\mathcal{S}$ is equivalent to an invariant function on $\mathcal{S}$. There are problems where Assumption A (ii) cannot be checked, since the only known theorem that covers Assumption A (ii) assumes a certain property of the structure of G, and, in addition can be applied only if $\mathcal{P}$ is a dominated family. This excludes application to nonparametric problems. Theorem 6.1, in Section 6, avoids those handicaps by using Assumption B, which is the existence of an invariant conditional probability distribution. Such a distribution is usually easily exhibited in cases where the conditional probability distribution is discrete, such as in certain nonparametric problems. Example 6.1 illustrates this case. On the other hand, in a large class of problems involving a family of densities with respect to Lebesgue measure, Assumption B may be difficult to verify directly. In order to cope with such cases, Theorem 7.1 in Section 7 gives sufficient conditions (called Assumption C) for Assumption B to be valid if $\mathcal{P}$ is a family of densities on n-space of the form $p_\theta(x) = g_\theta(s(x))h(x)$, where the various functions, g_θ, s, h, and each transformation g satisfy certain regularity conditions. This theorem could perhaps be regarded as a rigorized version of Cox's theorem [9]. Its advantage over Assumption A is that its conditions can be checked in a straightforward way. In particular, it does not involve the topological structure of G. The use of Theorem 7.1 is illustrated in examples 7.1, 7.2 and 7.3.

II.2. Preliminaries on transformations and conditional independence. Let $\mathfrak{X}$ be a space of points x, $\mathcal{A}$ a σ-field of subsets of $\mathfrak{X}$, P a probability measure on $\mathcal{A}$. If $\mathcal{A}_0 \subset \mathcal{A}$ and $\mathcal{A}_0$ is a σ-field, we shall simply denote this by $\mathcal{A}_0 \subset \mathcal{A}$, and we shall, for short, call $\mathcal{A}_0$ a subfield of $\mathcal{A}$, or simply a subfield. If $\mathcal{A}_1$ and $\mathcal{A}_2$ are two subfields, their intersection $\mathcal{A}_1 \cap \mathcal{A}_2$ is also a subfield. With the union $\mathcal{A}_1 \cup \mathcal{A}_2$ this is not usually the case. We shall denote by $\mathcal{A}_1 \vee \mathcal{A}_2$ the smallest subfield containing both $\mathcal{A}_1$ and $\mathcal{A}_2$. All functions will be understood to be $\mathcal{A}$-measurable real-

valued functions on $\mathfrak{X}$, and, if required in the context, P-integrable. If two functions f_1 and f_2 are equal except on a set of P-measure 0, i.e. $f_1 = f_2$ a.e. P, we shall denote this by $f_1 \sim f_2$, and sometimes term this "f_1 is equivalent to f_2." If $\mathcal{A}_0 \subset \mathcal{A}$ and f is $\mathcal{A}_0$-measurable, we shall sometimes term this "an $\mathcal{A}_0$ function f." The conditional expectation of f given $\mathcal{A}_0$, written $E(f \mid \mathcal{A}_0)$, is defined as any $\mathcal{A}_0$ function whose integral over any $A_0 \in \mathcal{A}_0$ equals the integral of f over A_0 (this follows Loève's definition [32]; Doob [11] relaxes the definition somewhat by including all functions that are equivalent to an $\mathcal{A}_0$ function with the above mentioned property). The conditional expectation of f given $\mathcal{A}_0$ is defined up to an equivalence within the $\mathcal{A}_0$ functions, and we shall sometimes speak of the various "versions" of this conditional expectation. The conditional probability of a set A, $P(A \mid \mathcal{A}_0)$, is the conditional expectation of the indicator of A.

Let g be a 1–1 transformation of $\mathfrak{X}$ onto a space $\mathcal{Y}$ of points y. We write $\mathcal{Y} = g\mathfrak{X}$, and $y = gx$ if y is the image of x. The point transformation g induces in a natural way a set transformation, which we shall also denote by g. Thus, gA is the image of $A \in \mathcal{A}$. The collection of all gA is obviously a σ-field of subsets of $\mathcal{Y}$, which we shall denote by $g\mathcal{A}$. Furthermore, g induces a probability distribution, denoted gP, on $g\mathcal{A}$: $gP(gA)$ is defined as $P(A)$. Finally, to each function f on $\mathfrak{X}$ corresponds a function on $\mathcal{Y}$, denoted gf: $gf(gx)$ is defined as $f(x)$. (This definition of gf makes sense even if f has an arbitrary range space.)

The transformation g produces an isomorphism between $(\mathfrak{X}, \mathcal{A}, P)$ and $(g\mathfrak{X}, g\mathcal{A}, gP)$. Thus, if $\mathcal{A}_0 \subset \mathcal{A}$ then $g\mathcal{A}_0 \subset g\mathcal{A}$, and if f is a P-integrable function on $\mathfrak{X}$ then gf is a gP-integrable function on $g\mathfrak{X}$ and a possible version of $E(gf \mid g\mathcal{A}_0)$ is $gE(f \mid \mathcal{A}_0)$, so that

$$E(gf \mid g\mathcal{A}_0) \sim gE(f \mid \mathcal{A}_0) \tag{2.1}$$

(the equivalence $\sim$ is here with respect to gP on $g\mathcal{A}_0$).

The considerations given so far will be applied in the case $\mathcal{Y} = \mathfrak{X}$, i.e. g is a 1-1 transformation of $\mathfrak{X}$ onto itself. In that case it makes sense to talk about the possibilities $g\mathcal{A} = \mathcal{A}$, $gf = f$, etc.

The rest of this section is devoted to propositions on conditional independence. Let $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ be three subfields; then $\mathcal{A}_1$ and $\mathcal{A}_2$ are defined in [32], p. 351, to be *conditionally independent given* $\mathcal{A}_3$ if for any $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$ we have $P(A_1A_2 \mid \mathcal{A}_3) \sim P(A_1 \mid \mathcal{A}_3)P(A_2 \mid \mathcal{A}_3)$. Instead of giving the definition in terms of conditional probabilities of sets, we may, equivalently, give it in terms of conditional expectations of integrable functions. Since this is more convenient in the sequel, we shall state

DEFINITION 2.1. Subfields $\mathcal{A}_1$ and $\mathcal{A}_2$ are conditionally independent given $\mathcal{A}_3$ if for any $\mathcal{A}_1$ function f_1 and $\mathcal{A}_2$ function f_2 we have

$$E(f_1f_2 \mid \mathcal{A}_3) \sim E(f_1 \mid \mathcal{A}_3)E(f_2 \mid \mathcal{A}_3). \tag{2.2}$$

The definition of unconditional independence of $\mathcal{A}_1$ and $\mathcal{A}_2$ follows by taking in (2.2) $\mathcal{A}_3 = \{\mathfrak{X}, \phi\}$ (i.e. $\mathcal{A}_3$ is the trivial subfield) and by replacing $\sim$ by $=$. The following theorem is proved in [32], p. 351.

THEOREM 2.1. *$\mathcal{A}_1$ and $\mathcal{A}_2$ are conditionally independent given $\mathcal{A}_3$ if and only if for every $\mathcal{A}_2$ function f_2,*

$$E(f_2 \mid \mathcal{A}_1 \vee \mathcal{A}_3) \sim E(f_2 \mid \mathcal{A}_3). \tag{2.3}$$

From Theorem 2.1 and the fact that $(\mathcal{A}_1 \vee \mathcal{A}_3) \vee \mathcal{A}_3 = \mathcal{A}_1 \vee \mathcal{A}_3$ follows

COROLLARY 2.1. *$\mathcal{A}_1$ and $\mathcal{A}_2$ are conditionally independent given $\mathcal{A}_3$ if and only if $\mathcal{A}_1 \vee \mathcal{A}_3$ and $\mathcal{A}_2$ are conditionally independent given $\mathcal{A}_3$.*

By taking in Theorem 2.1 $\mathcal{A}_3 = \{\mathfrak{X}, \phi\}$ we have

COROLLARY 2.2. *$\mathcal{A}_1$ and $\mathcal{A}_2$ are independent if and only if for every $\mathcal{A}_2$ function f_2 we have $E(f_2 \mid \mathcal{A}_1) \sim Ef_2$.*

If $\mathcal{A}_1$ and $\mathcal{A}_2$ are conditionally independent given $\mathcal{A}_3$, if $\mathcal{A}_0 \subset \mathcal{A}_1$ and if f_1 is $\mathcal{A}_0$-measurable, then f_1 is also $\mathcal{A}_3$-measurable so that (2.2) holds. We have therefore

LEMMA 2.1. *If $\mathcal{A}_1$ and $\mathcal{A}_2$ are conditionally independent given $\mathcal{A}_3$, and $\mathcal{A}_0 \subset \mathcal{A}_1$, then $\mathcal{A}_0$ and $\mathcal{A}_2$ are conditionally independent given $\mathcal{A}_3$.*

By taking $\mathcal{A}_3 = \{\mathfrak{X}, \phi\}$ in Lemma 2.1 we have

COROLLARY 2.3. *If $\mathcal{A}_1$ and $\mathcal{A}_2$ are independent and $\mathcal{A}_0 \subset \mathcal{A}_1$, then $\mathcal{A}_0$ and $\mathcal{A}_2$ are independent.*

LEMMA 2.2. *If $\mathcal{A}_1$ and $\mathcal{A}_2$ are independent and $\mathcal{A}_3 \subset \mathcal{A}_1$ then $\mathcal{A}_1$ and $\mathcal{A}_2$ are conditionally independent given $\mathcal{A}_3$.*

PROOF. Let f_2 be $\mathcal{A}_2$-measurable. Using Theorem 2.1 and observing $\mathcal{A}_1 \vee \mathcal{A}_3 = \mathcal{A}_1$, we have to show

$$E(f_2 \mid \mathcal{A}_1) \sim E(f_2 \mid \mathcal{A}_3). \tag{2.4}$$

This is true because both sides of (2.4) are $\sim Ef_2$. For the left hand side this follows from Corollary 2.2, and for the right hand side by first applying Corollary 2.3 with $\mathcal{A}_0 = \mathcal{A}_3$.

The various propositions on conditional independence in this section have their obvious analogues in terms of random variables. For instance, Corollary 2.1 would read: X and Y are conditionally independent given Z if and only if (X, Z) and Y are conditionally independent given Z. Lemma 2.2 would read: if X and Y are independent, h a function of X, then X and Y are conditionally independent given $h(X)$.

II. 3. Sufficiency and invariance in the nonsequential case. Assumption A. Let $\mathfrak{X}$ and $\mathcal{A}$ be as in Section 2, and let $\mathcal{P}$ be a family of probability measures P. If we write $f_1 \sim f_2$ this will mean now that $f_1 = f_2$ a.e. $\mathcal{P}$, i.e. the set on which the equality does not hold has P-measure 0 for every $P \in \mathcal{P}$. All functions are understood to be P-integrable for every $P \in \mathcal{P}$ whenever this is required in the context. The expectation with respect to P will now be written E_P.

If $\mathcal{A}_1$ and $\mathcal{A}_2$ are any subfields of $\mathcal{A}$, with $\mathcal{A}_2 \subset \mathcal{A}_1$, we say that *$\mathcal{A}_2$ is sufficient for $\mathcal{A}_1$ if for every $\mathcal{A}_1$ function f_1 there is an $\mathcal{A}_2$ function f_2 such that $E_P(f_1 \mid \mathcal{A}_2) \sim f_2$ for all $P \in \mathcal{P}$.* In particular, let $\mathcal{A}_S$ be sufficient for $\mathcal{A}$.

Let G be a group of transformations g of $\mathfrak{X}$ one–one onto itself, such that for each g,

(a) $g\mathfrak{A} = \mathfrak{A}$,

(b) $gP \in \mathcal{P}$ whenever $P \in \mathcal{P}$.

A function f is called *invariant* [30] if $gf = f$ for each $g \in G$; f is called *almost invariant* if $gf \sim f$ for each $g \in G$ (where the exceptional $\mathcal{P}$-null set may depend on g). A subset of $\mathfrak{X}$ is called invariant if its indicator is. The invariant members of $\mathfrak{A}$ form clearly a subfield. We shall denote it by $\mathfrak{A}_I$. Hence, for each $A \in \mathfrak{A}_I$ we have $gA = A$. The members of $\mathfrak{A}$ that are in both $\mathfrak{A}_S$ and $\mathfrak{A}_I$ constitute the subfield $\mathfrak{A}_{SI} = \mathfrak{A}_S \cap \mathfrak{A}_I$. Clearly $\mathfrak{A}_{SI}$ is a subfield both of $\mathfrak{A}_S$ and of $\mathfrak{A}_I$.

Concerning the relation between G, $\mathfrak{A}_S$ and $\mathfrak{A}_I$ we shall make the following assumption:

ASSUMPTION A. (i) $g\mathfrak{A}_S = \mathfrak{A}_S$ *for each* $g \in G$; (ii) *if* f_S *is* $\mathfrak{A}_S$*-measurable and almost invariant, there exists an* $\mathfrak{A}_{SI}$ *function* f_{SI} *such that* $f_{SI} \sim f_S$.

Before stating the main result in this section (conclusion of Theorem 3.1), it is convenient to state first the two following lemmas.

LEMMA 3.1. *Under Assumption A* (i), *if* f *is invariant then any version of* $E(f \mid \mathfrak{A}_S)$ *is almost invariant.*

PROOF. In (2.1) on the left hand side we have $gf = f$ since f is invariant, and, replacing $\mathfrak{A}_0$ by $\mathfrak{A}_S$, we have $g\mathfrak{A}_S = \mathfrak{A}_S$ by Assumption A (i). Thus, (2.1) reads $E(f \mid \mathfrak{A}_S) \sim gE(f \mid \mathfrak{A}_S)$. If f_S is any version of $E(f \mid \mathfrak{A}_S)$, we have $f_S \sim gf_S$, which is the conclusion of the lemma.

Using Lemma 3.1, and Assumption A (ii) we have immediately

LEMMA 3.2. *Under Assumption A, if* f *is invariant there exists an* $\mathfrak{A}_{SI}$ *function* f_{SI} *such that* $f_{SI} \sim E(f \mid \mathfrak{A}_S)$.

The following theorem was first stated and proved by Stein (unpublished) under slightly different assumptions. The analogue in terms of statistics is given in Part I, Section 2. The statement and proof given here follow consistently the language of subfields.

THEOREM 3.1. *Under Assumption A,* $\mathfrak{A}_{SI}$ *is sufficient for* $\mathfrak{A}_I$.

PROOF. We have to show that if f is $\mathfrak{A}_I$ measurable, there exists an $\mathfrak{A}_{SI}$ function f_{SI} such that

$$E_P(f \mid \mathfrak{A}_{SI}) \sim f_{SI} \qquad \text{for all } P \in \mathcal{P}. \tag{3.1}$$

To show this, let f_S be any version of $E(f \mid \mathfrak{A}_S)$. Since $\mathfrak{A}_{SI} \subset \mathfrak{A}_S$, we have by a well-known property of iterated conditional expectations ([11], p. 37):

$$E_P(f \mid \mathfrak{A}_{SI}) \sim E_P(f_S \mid \mathfrak{A}_{SI}), \qquad P \in \mathcal{P}. \tag{3.2}$$

From Lemma 3.2 we know that there is an $\mathfrak{A}_{SI}$ function f_{SI} such that $f_{SI} \sim f_S$. Substituting f_{SI} for f_S on the right hand side in (3.2), and observing $E_P(f_{SI} \mid \mathfrak{A}_{SI}) \sim f_{SI}$, we have (3.1). This concludes the proof.

There are a few additional properties of a subfield of interest besides sufficiency. One is *completeness*, another is *minimal sufficiency*. Are these properties inherited by $\mathfrak{A}_{SI}$ if valid for $\mathfrak{A}_S$? We recall [31] that $\mathcal{P}$ on $\mathfrak{A}_S$ is called *complete* if, for an $\mathfrak{A}_S$ function f, $E_P f = 0$ for all $P \in \mathcal{P}$ implies $f \sim 0$. It follows then immediately from the definition that if $\mathcal{P}$ is complete on $\mathfrak{A}_S$, it is also complete on any subfield

of $\mathfrak{A}_S$, in particular on $\mathfrak{A}_{SI}$. Hence, completeness is inherited by $\mathfrak{A}_{SI}$. The situation is different for minimal sufficiency. We recall that a subfield $\mathfrak{A}_2$, sufficient for $\mathfrak{A}_1$, is called *minimal sufficient* [31] (or *necessary and sufficient* in [4]) if every subfield sufficient for $\mathfrak{A}_1$ contains $\mathfrak{A}_2$, up to $\mathcal{P}$-null sets. The following counter example shows that if $\mathfrak{A}_S$ is minimal sufficient for $\mathfrak{A}$, it is not necessarily true that $\mathfrak{A}_{SI}$ is minimal sufficient for $\mathfrak{A}_I$. Let $X_1, \cdots, X_n$ be independently normal with common unknown standard deviation σ and common mean $c\sigma$, where $c \neq 0$ is a known real number. Let G consist of all transformations of the form $x_i \to gx_i$, $i = 1, \cdots, n$, where g is any positive number. Then $\mathfrak{A}_I$ is induced by the maximal invariant $(X_1/X_n, \cdots, X_{n-1}/X_n, \operatorname{sgn} X_n)$. Let $\bar{X}$ and S be the sample mean and standard deviation, respectively; then $\mathfrak{A}_S$ is induced by the minimal sufficient (but not complete) statistic $(\bar{X}, S)$, and $\mathfrak{A}_{SI}$ is induced by the statistic $\bar{X}/S$. However, $\mathfrak{A}_{SI}$ is not minimal sufficient for $\mathfrak{A}_I$ since the trivial subfield $\{\mathfrak{X}, \phi\}$ is contained in $\mathfrak{A}_{SI}$, not equivalent to it, and also sufficient for $\mathfrak{A}_I$. The latter is true because the distribution of the maximal invariant is free of σ, so that any $\mathfrak{A}_I$ function has a fixed distribution.

The relevance of the conclusion of Theorem 3.1 for testing problems is that for every invariant test function φ_I there exists an $\mathfrak{A}_{SI}$-measurable test function φ_{SI} with the same power function. If $\mathfrak{A}_S$ is complete, the property mentioned in the preceding sentence is possessed not only by the invariant tests, but by all test functions φ whose power function is invariant, including the φ_I as special cases. To see this we apply first Lemma 2, p. 227, in [30] to $E(\varphi \mid \mathfrak{A}_S)$, then Assumption A (ii), and conclude that there is an invariant version φ_{SI} of $E(\varphi \mid \mathfrak{A}_S)$. Under these circumstances, if a test enjoys a certain optimum property within the class of $\mathfrak{A}_{SI}$-measurable tests, it also enjoys this property among all tests whose power function is invariant. (An analogous statement may be made in the sequential case, replacing "power function" by "joint distribution of decision and sample size".)

We conclude this section by an interpretation of Theorem 3.1 in the language of conditional independence. The latter notion was defined in Section 2 in the case of one probability measure P. In the remainder of this paper we shall call $\mathfrak{A}_1$ and $\mathfrak{A}_2$ conditionally independent given $\mathfrak{A}_3$ if for every $P \in \mathcal{P}$ (2.2) holds, with E replaced by E_P.

LEMMA 3.3. *The following statements are equivalent:*

(i) *If f_I is invariant, there exists an $\mathfrak{A}_{SI}$ function f_{SI} such that $f_{SI} \sim E(f_I \mid \mathfrak{A}_S)$.*

(ii) *If f_I is invariant, then for every $P \in \mathcal{P}$,*

$$E(f_I \mid \mathfrak{A}_S) \sim E_P(f_I \mid \mathfrak{A}_{SI}). \tag{3.3}$$

(iii) *$\mathfrak{A}_S$ and $\mathfrak{A}_I$ are conditionally independent given $\mathfrak{A}_{SI}$.*

PROOF. (i) follows from (ii) immediately by taking f_{SI} to be any version of $E_P(f_I \mid \mathfrak{A}_{SI})$ for any P. Conversely, (ii) follows from (i) by writing the right hand side of (3.3) as $E_P(E(f_I \mid \mathfrak{A}_S) \mid \mathfrak{A}_{SI})$. Then (3.3) follows after remarking that both sides are equivalent to f_{SI}, using (i). The equivalence of (ii) and (iii) follows immediately from Theorem 2.1 by taking in this theorem $\mathfrak{A}_1 = \mathfrak{A}_S$, $\mathfrak{A}_2 = \mathfrak{A}_I$, $\mathfrak{A}_3 = \mathfrak{A}_{SI}$ (so that $\mathfrak{A}_1 \vee \mathfrak{A}_3 = \mathfrak{A}_S$), $f_2 = f_I$, and replacing E by E_P.

We recognize (i) of Lemma 3.3 as the conclusion of Lemma 3.2. Using Lemma 3.3 (i) and (iii) we see then that Lemma 3.2 is equivalent to

THEOREM 3.2. *Under Assumption A, $\mathcal{A}_S$ and $\mathcal{A}_I$ are conditionally independent given $\mathcal{A}_{SI}$.*

Since the conclusion of Theorem 3.1 followed from the conclusion of Lemma 3.2, it follows then also from the conclusion of Theorem 3.2. We could therefore interpret the results as follows: Assumption A is used to establish the conditional independence of $\mathcal{A}_S$ and $\mathcal{A}_I$ given $\mathcal{A}_{SI}$, and this in turn implies the sufficiency of $\mathcal{A}_{SI}$ for $\mathcal{A}_I$. (It was noticed by J. K. Ghosh that if $\mathcal{P}$ on $\mathcal{A}_S$ is complete, then the conclusions of Theorem 3.1 and 3.2 are equivalent.)

One of the advantages of Theorem 3.2 is that it is symmetric in $\mathcal{A}_S$ and $\mathcal{A}_I$. Therefore, any statement implied by the conclusion of Theorem 3.2 remains true if the subscripts S and I are interchanged. For example, if we do this in (3.3) (after replacing on the left hand side E by E_P) we get for every $\mathcal{A}_S$ function f_S and every P, $E_P(f_S \mid \mathcal{A}_I) \sim E_P(f_S \mid \mathcal{A}_{SI})$.

II.4. Sufficiency, invariance and transitivity in the sequential case. Let $\{\mathcal{A}_n, n \geqq 1\}$ be a sequence of subfields of $\mathcal{A}$, and let, for each $n \geqq 1$, $\mathcal{A}_{Sn}$, $\mathcal{A}_{In}$ and $\mathcal{A}_{SIn}$ be subfields of $\mathcal{A}_n$, defined in the same way as $\mathcal{A}_S$, $\mathcal{A}_I$ and $\mathcal{A}_{SI}$ were in Section 3. For each n, $\mathcal{A}_{Sn}$ is sufficient for $\mathcal{A}_n$. We shall express this by saying that $\{\mathcal{A}_{Sn}\}$ is a *sufficient sequence for* $\{\mathcal{A}_n\}$. From Theorem 3.1 we know that if Assumption A is valid for each n, then $\{\mathcal{A}_{SIn}\}$ is a sufficient sequence for $\{\mathcal{A}_{In}\}$. Besides the notion of sufficiency there is in the sequential case an additional notion, called *transitivity*, and introduced by Bahadur [4].

DEFINITION 4.1. Let $\{\mathcal{B}_n, n \geqq 1\}$ and $\{\mathcal{B}_{0n}, n \geqq 1\}$ be two sequences of subfields such that $\mathcal{B}_{0n} \subset \mathcal{B}_n$ for each n. The sequence $\{\mathcal{B}_{0n}\}$ is said to be a transitive sequence for $\{\mathcal{B}_n\}$ if for every n, every $\mathcal{B}_{0(n+1)}$ function f and every $P \in \mathcal{P}$ we have

$$E_P(f \mid \mathcal{B}_n) \sim E_P(f \mid \mathcal{B}_{0n}). \tag{4.1}$$

The importance of $\{\mathcal{A}_{Sn}\}$ being a sufficient and transitive sequence for $\{\mathcal{A}_n\}$ has been pointed out in [4]. A discussion can also be found in Part I, Section 4. This section will be concerned mainly with the question of transitivity. It is of some interest that Definition 4.1 is equivalent to a statement in terms of conditional independence of subfields, as follows:

Theorem 4.1. *$\{\mathcal{B}_{0n}\}$ is a transitive sequence for $\{\mathcal{B}_n\}$ if and only if for each $n \geqq 1$ $\mathcal{B}_n$ and $\mathcal{B}_{0(n+1)}$ are conditionally independent given $\mathcal{B}_{0n}$.*

PROOF. Let f be $\mathcal{B}_{0(n+1)}$-measurable. Apply Theorem 2.1 with E replaced by E_P, $\mathcal{A}_1 = \mathcal{B}_n$, $\mathcal{A}_2 = \mathcal{B}_{0(n+1)}$, $\mathcal{A}_3 = \mathcal{B}_{0n}$ (so that $\mathcal{A}_1 \vee \mathcal{A}_3 = \mathcal{B}_n$) and $f_2 = f$. Then Theorem 2.1 states that $\mathcal{B}_n$ and $\mathcal{B}_{0(n+1)}$ are conditionally independent given $\mathcal{B}_{0n}$ if and only if for each P (4.1) holds.

Two questions will be investigated in the remainder of this section. The first is whether $\{\mathcal{A}_{SIn}\}$ is a transitive sequence for $\{\mathcal{A}_{Sn}\}$ if $\{\mathcal{A}_{Sn}\}$ is a transitive sequence for $\{\mathcal{A}_n\}$. This question was answered by Ghosh [16], Theorem 2, Chapter 4,

in the affirmative under slightly more restrictive conditions than we shall impose in Theorem 4.2. The second question is under what conditions $\{\mathcal{A}_{2n}\}$ is a transitive sequence for $\{\mathcal{A}_n\}$. This question was suggested by Ghosh's theorem quoted above, and the answer, as formulated in Theorem 4.3 below, is essentially contained in the proof of Theorem 11.5 in [4]. Our proof of Theorem 4.3 is entirely in terms of conditional independence of subfields. Note that Theorems 4.2 and 4.3 are detached from sufficiency and invariance considerations.

THEOREM 4.2. *For each $n \geq 1$, let $\mathcal{A}_n$, $\mathcal{A}_{1n}$, $\mathcal{A}_{2n}$ and $\mathcal{A}_{3n}$ be subfields, with $\mathcal{A}_{1n} \subset \mathcal{A}_n$, $\mathcal{A}_{2n} \subset \mathcal{A}_n$ and $\mathcal{A}_{3n} \subset \mathcal{A}_{1n} \cap \mathcal{A}_{2n}$, such that $\mathcal{A}_{1n}$ and $\mathcal{A}_{2n}$ are conditionally independent given $\mathcal{A}_{3n}$. Then if $\{\mathcal{A}_{1n}\}$ is a transitive sequence for $\{\mathcal{A}_n\}$, $\{\mathcal{A}_{3n}\}$ is a transitive sequence for $\{\mathcal{A}_{2n}\}$.*

PROOF. We have to show that if f is $\mathcal{A}_{3(n-1)}$-measurable, then for each P

$$E_P(f \mid \mathcal{A}_{2n}) \sim E_P(f \mid \mathcal{A}_{3n}). \tag{4.2}$$

Because f is $\mathcal{A}_{3(n-1)}$-measurable, and $\mathcal{A}_{3(n-1)} \subset \mathcal{A}_{1(n-1)}$, f is also $\mathcal{A}_{1(n-1)}$-measurable. Since by assumption $\{\mathcal{A}_{1n}\}$ is a transitive sequence for $\{\mathcal{A}_n\}$, we have for each P,

$$E_P(f \mid \mathcal{A}_n) \sim E_P(f \mid \mathcal{A}_{1n}). \tag{4.3}$$

We apply now Theorem 2.1, replacing in (2.3) E by E_P, $\mathcal{A}_1$ by $\mathcal{A}_{2n}$, $\mathcal{A}_2$ by $\mathcal{A}_{1n}$, $\mathcal{A}_3$ by $\mathcal{A}_{3n}$, f_2 by $E_P(f \mid \mathcal{A}_{1n})$. Then $\mathcal{A}_1 \vee \mathcal{A}_3$ is replaced by $\mathcal{A}_{2n} \vee \mathcal{A}_{3n} = \mathcal{A}_{2n}$, and (2.3) reads

$$E_P(E_P(f \mid \mathcal{A}_{1n}) \mid \mathcal{A}_{2n}) \sim E_P(E_P(f \mid \mathcal{A}_{1n}) \mid \mathcal{A}_{3n}). \tag{4.4}$$

We have now

$$\begin{aligned} E_P(f \mid \mathcal{A}_{2n}) &\sim E_P(E_P(f \mid \mathcal{A}_n) \mid \mathcal{A}_{2n}) \\ &\sim E_P(E_P(f \mid \mathcal{A}_{1n}) \mid \mathcal{A}_{2n}) \qquad \text{by (4.3)} \\ &\sim E_P(E_P(f \mid \mathcal{A}_{1n}) \mid \mathcal{A}_{3n}) \qquad \text{by (4.4)} \\ &\sim E_P(f \mid \mathcal{A}_{3n}) \end{aligned}$$

which is (4.2).

We are especially interested in applying Theorem 4.2 to the case $\mathcal{A}_{3n} = \mathcal{A}_{1n} \cap \mathcal{A}_{2n}$. Taking in Theorem 4.2 $\mathcal{A}_{1n} = \mathcal{A}_{5n}$, $\mathcal{A}_{2n} = \mathcal{A}_{4n}$, $\mathcal{A}_{3n} = \mathcal{A}_{6n}$, we have

COROLLARY 4.1. *If the conclusion of Theorem 3.2 is valid for each $n \geq 1$, and if $\{\mathcal{A}_{5n}\}$ is a sufficient and transitive sequence for $\{\mathcal{A}_n\}$, then $\{\mathcal{A}_{6n}\}$ is a sufficient and transitive sequence for $\{\mathcal{A}_{4n}\}$.*

THEOREM 4.3 *Suppose a sequence $\mathcal{B}_1, \mathcal{B}_2, \cdots$ of independent subfields of $\mathcal{A}$ is given; suppose $\mathcal{A}_n = \mathcal{B}_1 \vee \cdots \vee \mathcal{B}_n$; let $\{\mathcal{A}_{0n}\}$ be given such that for each n $\mathcal{A}_{0n} \subset \mathcal{A}_n$ and $\mathcal{A}_{0(n+1)} \subset \mathcal{A}_{0n} \vee \mathcal{B}_{n+1}$; then $\{\mathcal{A}_{0n}\}$ is a transitive sequence for $\{\mathcal{A}_n\}$.*

PROOF. By the construction of $\{\mathcal{A}_n\}$, $\mathcal{A}_n$ and $\mathcal{B}_{n+1}$ are independent. We apply Lemma 2.2 with $\mathcal{A}_1 = \mathcal{A}_n$, $\mathcal{A}_2 = \mathcal{B}_{n+1}$, $\mathcal{A}_3 = \mathcal{A}_{0n}$ and conclude that $\mathcal{A}_n$ and $\mathcal{B}_{n+1}$ are conditionally independent given $\mathcal{A}_{0n}$. Applying Corollary 2.1 we have that $\mathcal{A}_n$ and $\mathcal{B}_{n+1} \vee \mathcal{A}_{0n}$ are conditionally independent given $\mathcal{A}_{0n}$. We apply now Lemma 2.1 with $\mathcal{A}_1 = \mathcal{B}_{n+1} \vee \mathcal{A}_{0n}$, $\mathcal{A}_0 = \mathcal{A}_{0(n-1)}$, $\mathcal{A}_2 = \mathcal{A}_n$, $\mathcal{A}_3 = \mathcal{A}_{0n}$, and

conclude that $\mathcal{A}_{s(n+1)}$ and $\mathcal{A}_n$ are conditionally independent given $\mathcal{A}_{sn}$. The desired result now follows from Theorem 4.1 with $\mathcal{B}$ replaced by $\mathcal{A}$.

When sampling from an exponential family of distributions the assumptions of Theorem 4.3 usually apply. For instance, if $X_1, X_2, \cdots$ are independent and identically distributed according to a normal distribution with unknown mean and known variance, then $\mathcal{B}_n$ is induced by X_n, $\mathcal{A}_n$ by $(X_1, \cdots, X_n)$, and $\mathcal{A}_{sn}$ by $T_n = X_1 + \cdots + X_n$. Hence $\mathcal{A}_{s(n+1)}$, which is induced by $T_n + X_{n+1}$, is a subfield of $\mathcal{A}_{sn} \vee \mathcal{B}_{n+1}$. Other examples of the use of Theorem 4.3 are given in Part I.

II.5. Discussion of Assumption A. As explained in Section 1, Assumption A (i) is a very natural one to make. If $\mathcal{A}_s$ is induced by a sufficient statistic s with range $\mathcal{S}$, then Assumption A (i) is closely related to the property that every $g \in G$ induces a 1–1 transformation of $\mathcal{S}$ onto itself. The following theorem is due to C. M. Stein (unpublished) and gives conditions under which Assumption A (i) holds.

THEOREM 5.1 (Stein). *If $\mathcal{A}_s$ is minimal sufficient for $\mathcal{A}$, and if $\mathcal{A}_s$ contains all $\mathcal{P}$-null sets, then Assumption A* (i) *is valid.*

PROOF. For each $g \in G$, due to the isomorphism described in Section 2 between $(\mathcal{X}, \mathcal{A}, P)$ and $(g\mathcal{X}, g\mathcal{A}, gP)$, we have that $g\mathcal{A}_s$ is minimal sufficient for $g\mathcal{P}$ on $g\mathcal{A}$. But $g\mathcal{P} = \mathcal{P}$ and $g\mathcal{A} = \mathcal{A}$, so that both $\mathcal{A}_s$ and $g\mathcal{A}_s$ are minimal sufficient for $\mathcal{P}$ on $\mathcal{A}$. Since they both contain all $\mathcal{P}$-null sets, they must be the same.

It should be remarked that Assumption A (i) is usually easy to check directly, and has been found to hold in all interesting examples, whereas it is often not true that $\mathcal{A}_s$ contains all $\mathcal{P}$ null sets, in which case Theorem 5.1 is not applicable.

Assumption A can be phrased in the following way. Noting that $g\mathcal{A}_s = \mathcal{A}_s$ by Assumption A (i), we can consider $\mathcal{A}_s$ as our basic σ-field, instead of $\mathcal{A}$, i.e. consider only $\mathcal{A}_s$-measurable functions. Assumption A (ii) then says that every almost invariant function is equivalent to an invariant function. In applications $\mathcal{A}_s$ is often induced by a statistic s, and G induces a group of transformations on the range $\mathcal{S}$ of s. Considering then only measurable functions on $\mathcal{S}$, and invariance relative to the induced group of transformations on $\mathcal{S}$, Assumption A (ii) again says that every almost invariant function is equivalent to an invariant function. This assumption holds in a good many cases, as implied by a theorem of Lehmann [30], p. 225. However, Lehmann's theorem cannot be applied unless $\mathcal{P}$ is dominated, which excludes many interesting nonparametric cases. Furthermore, Lehmann's theorem requires the existence of a σ-finite measure on G possessing a certain invariance property, so that the applicability depends rather heavily on the topological structure of G. In nonparametric examples the group G is usually of such a nature that it is not known how to verify the existence of a measure with the desired properties.

In some problems G is finite. In that case, and, more generally, when G is countable, Assumption A (ii) is automatically fulfilled.

II.6. Assumption B: invariant conditional probability distribution. The dis-

cussion in Section 5 brings out the desirability of having another assumption, alternative to Assumption A, that also permits the conclusions of Theorems 3.1 and 3.2. The following Assumption B achieves this aim, by introducing a function Q, which we shall call an *invariant conditional probability distribution*. If Q satisfies merely (i) and (ii) of Assumption B, it has been called a conditional probability distribution by Doob [11] (except that in [11] the measurability condition is slightly weaker), and a regular conditional probability by Loève [32]. (We use here the symbol Q instead of the more customary P, since there is only one conditional probability distribution due to the sufficiency of $\mathfrak{A}_s$, whereas there is a whole family $\mathcal{P}$ of distributions P.) The additional condition (iii) gives Q a certain invariance property.

ASSUMPTION B. *There is a set $A_{SI} \subset \mathfrak{A}_{SI}$ of $\mathcal{P}$-measure* 1, *and a real valued function Q on $\mathfrak{A} \times \mathfrak{X}$, with $Q(A, x) = 0$ for every $A \in \mathfrak{A}$, $x \in A_{SI}$, such that*

(i) *for every $x \in A_{SI}$, $Q(\cdot, x)$ is a probability distribution on $\mathfrak{A}$;*

(ii) *for every $A \in \mathfrak{A}$, $Q(A, \cdot)$ is a version of $P(A \mid \mathfrak{A}_s)$;*

(iii) *for every $x \in X$, $A \in \mathfrak{A}$ and $g \in G$, $Q(gA, gx) = Q(A, x)$.*

The reason we have to distinguish in Assumption B between $\mathfrak{X}$ and A_{SI} is that it is not always possible in applications to satisfy (i) for all $x \in \mathfrak{X}$, and at the same time satisfy (iii). Regarding condition (iii), it is merely necessary to verify this for $x \in A_{SI}$, since if $x \notin A_{SI}$ then also $gx \notin A_{SI}$; so that by definition $Q(gS, gx) = Q(A, x) = 0$.

Strictly speaking, (ii) and (iii) of Assumption B would be sufficient to obtain the conclusions of Theorems 3.1 and 3.2, since for any invariant set A (ii) and (iii) say that there is an invariant version of $P(A \mid \mathfrak{A}_s)$. However, it is more natural to make (i) also part of Assumption B since the purpose of this assumption is to apply it to cases where a conditional probability distribution satisfying (i) and (ii) is readily exhibited, and where (iii) can then subsequently be verified. It should also be remarked at this point that seemingly Assumption B does not contain or imply Assumption A (i), even though the latter was announced in Section 1 to be assumed throughout this paper. It is true that Assumption A (i) is not needed for Theorem 6.1 below, but it is equally true that essentially Assumption A (i) is implied by Assumption B. To see this, consider the subfield generated by the totality of functions $Q(A, \cdot)$, for $A \in \mathfrak{A}$. This subfield can easily be checked to be sufficient, contained in $\mathfrak{A}_s$, differing from $\mathfrak{A}_s$ only in null sets, and satisfying Assumption A (i). Thus, if Assumption A (i) is not satisfied by $\mathfrak{A}_s$, the latter can be replaced by an equivalent subfield that does satisfy it.

Doob [11] has shown (a proof is also in [32]) that a possible version of $E(f \mid \mathfrak{A}_s)$ is defined by

$$E(f \mid \mathfrak{A}_s)(x) = \int f(x')Q(dx', x). \tag{6.1}$$

If f is invariant, and Q satisfies Assumption B, then $E(f \mid \mathfrak{A}_s)$ as defined by (6.1) can immediately be checked to be invariant. Since $E(f \mid \mathfrak{A}_s)$ is also $\mathfrak{A}_s$-measurable, it follows that the version of $E(f \mid \mathfrak{A}_s)$ given by (6.1) is an $\mathfrak{A}_{SI}$ function. This is precisely (i) of Lemma 3.3, which is equivalent to (iii) of the

same lemma, i.e. the conclusion of Theorem 3.2. We have therefore proved

THEOREM 6.1. *If Assumption B holds, then the conclusions of Theorems* 3.1 *and* 3.2 *are valid.*

We shall now give a nonparametric example of the use of Theorem 6.1.

EXAMPLE 6.1. Let $X_1, \cdots, X_n$ be independent random variables with common unknown distribution. Let the group G consist of transformations g defined by $gx = (h(x_1), \cdots, h(x_n))$, where h is strictly monotonic, continuous, and maps the real line onto itself. A sufficient statistic is the unordered set $\{X_1, \cdots, X_n\}$. If π stands generically for a permutation of the coordinates of a point x, then $\mathcal{A}_0$ can be described as the family of those sets $A \subset \mathcal{X}$ that have the property: $x \in A \Rightarrow \pi x \in A$ for every π. For any x, let $m(x)$ be the number of distinct points of the form πx (if the coordinates of x are distinct, $m(x) = n!$). Given x and a set $A \subset \mathcal{X}$, let $m_A(x)$ be the number of distinct points πx that are in A. Then define $Q(A, x) = m_A(x)/m(x)$. One can verify immediately that Q satisfies Assumption B, with $A_{SI} = \mathcal{X} = n$-space.

II.7. Assumption C: invariant conditional density. Assumption B is less easy to verify if $Q(\cdot, x)$ is not a discrete probability distribution. Theorem 7.1 in this section is designed to cope with the continuous case. More specifically, it gives sufficient conditions for Assumption B to hold. These conditions will be called Assumption C. Assumption C is usually very easy to verify, and Theorem 7.1 may therefore be used as an alternative to Theorems 3.1 and 3.2 in those cases where the latter also apply. Two such cases will be illustrated in Examples 7.1 and 7.2. The normal theory examples in Part I may also be treated with Theorem 7.1. The real advantage of this theorem lies in the fact that Assumption C does not involve the topological structure of the group G, so that the theorem may be used in cases where we don't know how to verify Assumption A (ii). Example 7.3 will illustrate such a case, in which the family of distributions is nonparametric, and the conditional probability distribution continuous.

We shall precede Theorem 7.1 by

LEMMA 7.1. *Suppose F is an open subset of n-space, G a group of transformations of F onto itself, s a differentiable function from F into k-space ($k < n$) with range $\mathcal{S}$. Let $D(x)$ be the $n \times k$ matrix whose ij element is $\partial s_j/\partial x_i$ evaluated at $x \in F$. We make the following assumptions:*

(i) *for each $g \in G$, the transformation $x \to gx$ is continuously differentiable;*

(ii) *for each $g \in G$, $s(x) = s(x')$ implies $s(gx) = s(gx')$;*

(iii) *$D(x)$ is continuous and of rank k for each $x \in F$.*

Then each $g \in G$ induces a 1–1 *and bi-continuously differentiable transformation $\bar{g}$ of $\mathcal{S}$ onto itself, where for $s' \in \mathcal{S}$ we define $\bar{g}s' = s(gx)$ for any x such that $s' = s(x)$.*

PROOF. That the transformation g is well-defined follows from (ii). The transformation is 1–1 since if for some x, x', $\bar{g}s(x) = \bar{g}s(x')$, i.e. $s(gx) = s(gx')$, then $s(x) = s(x')$, using (ii) with g^{-1}. The transformations $\bar{g}$ obviously form a group $\bar{G}$, which is a homomorphism of G.

To show that $\bar{g}$ is bi-continuously differentiable (i.e. $\bar{g}$ and $\bar{g}^{-1}$ are continuously

differentiable), it is sufficient to show that $\bar{g}$ is differentiable, since the same conclusion will then apply to $\bar{g}^{-1}$. Incidentally, this will show that the Jacobian of the transformation $\bar{g}$ is everywhere positive on $s(F)$. In the following, points in Euclidean space will be denoted by row vectors, and the same notation will be used for vector-valued functions. Furthermore, g will be an arbitrary but fixed element of G. Let s_0 be an arbitrary point in $\mathcal{S}$, and let $x_0 \in F$ be such that $s(x_0) = s_0$, so that $\bar{g}s_0 = s(gx_0)$. It follows from (iii) that there is an $n \times (n-k)$ matrix E_0 such that the $n \times n$ matrix $(D(x_0), E_0)$ is nonsingular. Since $D(x)$ is continuous at x_0, $(D(x), E_0)$ is nonsingular in a neighborhood N_0' of x_0. Define the function u_0 from N_0' into $(n-k)$-space by $u_0 = xE_0$, and define $v_0 = (s, u_0)$, so that v_0 is differentiable, with matrix of partial derivatives $(D(x), E_0)$ nonsingular on N_0'. By an implicit function theorem there is then a neighborhood N_0 of x_0, $N_0 \subset N_0'$, such that on N_0 the function v_0 is 1-1 and bi-continuously differentiable. Let $M_0 = v_0(N_0)$; then v_0 is a 1–1 bi-continuously differentiable map of N_0 onto M_0. Similarly, there is a neighborhood N_1 of gx_0, and a function u_1 on N_1, such that $v_1 = (s, u_1)$ is a 1–1 bi-continuously differentiable map of N_1 onto $M_1 = v_1(N_1)$. Without loss of generality we may assume $N_1 = gN_0$. By (i) the transformation g maps N_0 continuously differentiably onto N_1. Let w be the composition of the three functions v_0^{-1}, g and v_1; then w maps M_0 onto M_1 continuously differentiably (actually bi-continuously differentiably). Write w_1 for the first k components of w, so that w_1 maps M_0 continuously differentiably onto $s(N_1)$. By the construction of w, any point $(s', u) \in M_0$ is mapped by w_1 into $\bar{g}s'$. Hence $\bar{g}s'$ is a continuously differentiable function of (s', u). But we know that $\bar{g}s'$ depends only on s'; hence $\bar{g}s'$ is a continuously differentiable function of s', for s' in a neighborhood of s_0. This concludes the proof of the lemma.

Although not needed here, we shall give without proof an explicit expression for the matrix of partial derivatives of the transformation $\bar{g}$, i.e. the matrix $W_1(s')$ whose ij element is $\partial(\bar{g}s')_j/\partial s_i'$, for $s' \in \mathcal{S}$. Let $G_0(x)$ be the matrix whose ij element is $\partial(gx)_j/\partial x_i$, let $D(x)$, s_0 and x_0 be as in Lemma 7.1, and put $D_0 = D(x_0)$, $D_1 = D(gx_0)$, $G_0 = G_0(x_0)$. Then $W_1(s_0) = (D_0'D_0)^{-1}D_0'G_0D_1$.

In the following we shall write g instead of $\bar{g}$, in conformity with the notation in Sections 2 and 3.

Assumption C. *$\mathfrak{X}$ is an n-dimensional Borel set, $\mathcal{A}$ the Borel subsets of $\mathfrak{X}$, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ with Θ an arbitrary index set, and with respect to n-dimensional Lebesgue measure P_θ has a density*

$$p_\theta(x) = g_\theta(s(x))h(x), \qquad x \in \mathfrak{X}, \tag{7.1}$$

in which s is a measurable function from $\mathfrak{X}$ into k-space ($k < n$) with range $\mathcal{S}$, g_θ and h are positive, real-valued measurable functions on $\mathcal{S}$, $\mathfrak{X}$, respectively, and s and h satisfy the conditions below. Let G, $\mathcal{A}_\mathcal{S}$, $\mathcal{A}_\mathcal{I}$ and $\mathcal{A}_{\mathcal{SI}}$ be as in Section 3 and suppose that there is an open set $A_{\mathcal{SI}} \in \mathcal{A}_{\mathcal{SI}}$ of $\mathcal{P}$-measure 1, such that on $A_{\mathcal{SI}}$:

(i) *for each $g \in G$ the transformation $x \to gx$ is continuously differentiable, and the Jacobian depends only on $s(x)$;*

(ii) *for each $g \in G$, $s(x) = s(x')$ implies $s(gx) = s(gx')$;*

(iii) *s is continuously differentiable, and the matrix $D(x)$, whose ij element is $\partial s_j/\partial x_i$, is of rank k;*

(iv) *for each $g \in G$, $h(gx)/h(x)$ depends only on $s(x)$.*

Note that Assumption C (ii) implies Assumption A (i).

THEOREM 7.1. *Assumption C implies Assumption B and therefore the conclusions of Theorems 3.1 and 3.2.*

PROOF. As in the proof of Lemma 7.1, to each $x \in A_{SI}$ we can assign a neighborhood and a 1–1 bi-continuously differentiable function on this neighborhood into n-space whose first k components coincide with s. Since A_{SI} is separable, we can cover it with a countable subfamily of these neighborhoods, and from this subfamily we may construct a family $\{N_\alpha, \alpha = 1, 2, \cdots\}$ of disjoint sets whose union is A_{SI} (the N_α are not necessarily open, but each N_α contains an open set). On each N_α we have then a function u_α into $(n - k)$-space, such that $v_\alpha = (s, u_\alpha)$ maps N_α 1–1 bi-continuously differentiably into n-space. Let J_α be the Jacobian of v_α^{-1} and define the real-valued function $h[v_\alpha]$ on $v_\alpha(N_\alpha)$ by

$$h[v_\alpha](s', u) = h(v_\alpha^{-1}(s', u))J_\alpha(s', u), \qquad (s', u) \in v_\alpha(N_\alpha). \tag{7.2}$$

Note that $h[v_\alpha] > 0$ for each α. In the following, the indicator of any set B will be denoted by $I[B]$, and the probability with respect to P_θ of B by $P_\theta B$. For any $A \in \mathcal{A}$ and $s' \in \mathcal{S}$ we put

$$K[A](s') = \sum_\alpha \int I[V_\alpha(N_\alpha A)](s', u)h[v_\alpha](s', u)\, du \tag{7.3}$$

and for $K[\mathfrak{X}](s')$ we shall simply write $K(s')$. Now we define, for $A \in \mathcal{A}$, $x \in A_{SI}$:

$$Q(A, x) = K[A](s')/K(s'), \qquad s' = s(x). \tag{7.4}$$

We shall assume for the time being that $K(s')$ is neither 0 nor ∞, and return to a discussion of the possibilities $K(s') = 0$ or ∞ later. Note that $K[A](s')$ does not change if on any N_α the function u_α is changed to u_α', such that (s, u_α') is again a 1–1 bi-continuously differentiable function on N_α. This remark also can be used to show that $K[A](s')$ does not depend on the particular choice of the family $\{N_\alpha\}$: if $\{N_\beta', \beta = 1, 2, \cdots\}$ is another choice, with 1–1 bi-continuously differentiable function (s, u_β') on N_β', then we may employ the family $\{N_\alpha N_\beta'\}$ and on $N_\alpha N_\beta'$ we may take either the function (s, u_α) or (s, u_β'), giving the same contribution to the double series defining $K[A](s')$.

We shall show now that Q defined in (7.4) satisfies (i), (ii) and (iii) of Assumption B. That $Q(\cdot, x)$ is a probability distribution for each $x \in A_{SI}$ is immediate, so that (i) is true. To show (ii) we first remark that each term on the right hand side in (7.3) is a measurable function of s', so that $Q(A, \cdot)$ is $\mathcal{A}_S$-measurable. Now let $B_0 \in \mathcal{A}_S$, then

$$\begin{aligned} P_\theta A B_0 &= \sum_\alpha P_\theta N_\alpha A B_0 \\ &= \sum_\alpha \int I[N_\alpha A B_0](x)\varrho_\theta(s(x))h(x)\, dx \\ &= \sum_\alpha \int \varrho_\theta(s')\, ds' \int I[v_\alpha(N_\alpha A B_0)](s', u)h[v_\alpha](s', u)\, du. \end{aligned} \tag{7.5}$$

It is readily verified that $I[v_\alpha(N_\alpha A B_0)] = I[v_\alpha(N_\alpha A)]I[s(B_0)]$. Therefore, we

obtain from (7.5):

$$(7.6)\qquad P_\beta A B_0 = \sum_\alpha \int_{s(B_0)} g_\beta(s')\,ds' \int I[v_\alpha(N_\alpha A)](s', u)h[v_\alpha](s', u)\,du$$
$$= \int_{s(B_0)} g_\beta(s')K[A](s')\,ds'$$

using (7.3). By taking in (7.6) $A = \mathfrak{X}$ we get

$$(7.7)\qquad P_\beta B_0 = \int_{s(B_0)} g_\beta(s')K(s')\,ds'.$$

It follows from (7.7) that the random variable $s(X)$ has a density p_β^s with respect to k-dimensional Lebesgue measure, given by

$$(7.8)\qquad p_\beta^s(s') = g_\beta(s')K(s').$$

We still have to show that the integral of $Q(A, \cdot)$ over B_0 with respect to P_β equals $P_\beta A B_0$. Now, using (7.4), we may compute this integral as

$$(7.9)\qquad \int_{s(B_0)} [K[A](s')/K(s')]p_\beta^s(s')\,ds'$$

and substituting (7.8) into (7.9) we obtain the right hand side of (7.6). This concludes the verification of (ii) of Assumption B. Next we shall verify (iii).

From (7.4) we have $Q(gA, gx) = K[gA](gs')/K(gs')$, and we have to show that this equals $Q(A, x)$. Put $N_\beta' = gN_\beta$, $\beta = 1, 2, \cdots$, and let u_β' be a function on N_β' into $(n-k)$-space, defined by

$$(7.10)\qquad u_\beta'(x) = u_\beta(g^{-1}x), \qquad x \in N_\beta'$$

(i.e. $u_\beta' = gu_\beta$). Putting $v_\beta' = (s, u_\beta')$ and using (i), it can easily be checked that v_β' maps N_β' 1-1 bi-continuously differentiably into n-space. Since $\{N_\beta', \beta = 1, 2, \cdots\}$ is a family of disjoint sets, covering A_{SI}, we have, for any $B \subset A_{SI}$, $B = \bigcup_\beta N_\beta' B$, and the sets of this union are disjoint. Applying this to $B = N_\alpha g A$ in the expression for $K[gA](gs')$, we have $I[v_\alpha(N_\alpha g A)] = \sum_\beta I[v_\alpha(N_\beta' N_\alpha g A)]$ so that

$$(7.11)\qquad K[gA](gs') = \sum_{\alpha\beta} \int I[v_\alpha(N_\beta' N_\alpha g A)](gs', u)h[v_\alpha](gs', u)\,du.$$

On $N_\beta' N_\alpha$ we shall use now v_β' instead of v_α. By a previous remark this does not change the contribution to the $\alpha\beta$th term on the right hand side of (7.11). Thus:

$$(7.12)\qquad K[gA](gs') = \sum_{\alpha\beta} \int I[v_\beta'(N_\beta' N_\alpha g A)](gs', u)h[v_\beta'](gs', u)\,du.$$

By virtue of the construction of v_β' one can easily check

$$(7.13)\qquad I[v_\beta'(N_\beta' N_\alpha g A)](gs', u) = I[v_\beta(N_\beta A g^{-1} N_\alpha)](s', u).$$

Substituting (7.13) into (7.12) and summing over α yields

$$(7.14)\qquad K[gA](gs') = \sum_\beta \int I[v_\beta(N_\beta A)](s', u)h[v_\beta'](gs', u)\,du.$$

Using (iv) of Assumption C, let $h(gx)/h(x) = c_1(s')$, where $s' = s(x)$. Using (i) of Assumption C, let the Jacobian $|\partial(gx)/\partial x|$ be $c_2(s')$. Finally, using Lemma 7.1, let the Jacobian $|\partial s'/\partial(gs')|$ be $c_3(s')$. All three c's are >0 on A_{SI}, so their product $c(s') = c_1(s')c_2(s')c_3(s')$ is also positive on A_{SI}. With help of (7.2) one

can verify then that

$$h[v_g'](gs', u) = c(s')h[v_g](s', u). \tag{7.15}$$

If we substitute (7.15) into (7.14) and compare the result with (7.3) we see that

$$K[gA](gs') = c(s')K[A](s') \tag{7.16}$$

where, as remarked before, $c(s') > 0$. Applying (7.16) to $A = \mathfrak{X}$ gives

$$K(gs') = c(s')K(s'). \tag{7.17}$$

Taking the ratio of (7.16) and (7.17), and using (7.4), gives the desired result $Q(gA, gx) = Q(A, x)$, demonstrating the validity of (iii) of Assumption B.

We return now to the question whether $K(s')$, in the denominator of (7.4), can be 0 or ∞. Since $g_\theta > 0$ on $\mathcal{S}$, we see from (7.8) that the subset A_{S0} of A_{SI} on which $K(s(x)) = 0$ or ∞ is of $\mathcal{P}$-measure 0 (actually one can easily show that $K(s')$ cannot be 0 on $s(A_{SI})$, but we shall not need this fact). Moreover, it follows from (7.17) that A_{S0} is an invariant set. The invariant conditional probability distribution Q, given by (7.4), is well-defined on $A_{SI} - A_{S0}$, which is an $\mathfrak{A}_{SI}$ set of $\mathcal{P}$-measure 1, so that Theorem 7.1 can be applied with $A_{SI} - A_{S0}$ instead of A_{SI}. This concludes the proof of the theorem.

Example 7.1. Let $X_1, \cdots, X_n (n \geq 3)$ be independent and identically distributed according to a normal distribution with mean μ, standard deviation σ, both unknown. Then we may take $\theta = (\mu, \sigma)$, $s = (s_1, s_2)$ with $s_1(x) = \sum x_i$, $s_2(x) = \sum x_i^2$, $h(x) = 1$, and

$$g_\theta(s) = ((2\pi)^{\frac{1}{2}}\sigma)^{-n} \exp[-(2\sigma^2)^{-1}s_2 + (\mu/\sigma^2)s_1 - (n\mu^2/2\sigma^2)].$$

Let G be the totality of transformations $x \to cx$, $c > 0$. The matrix $D(x)$ of Assumption C is of rank 2 unless all components of x are equal, i.e. unless x is on the equiangular line. The latter is of Lebesgue measure 0 and in $\mathfrak{A}_{SI}$, so that its complement can be taken as the set A_{SI} in Assumption C. All assumptions are easily verified to hold. As a maximal invariant statistic based on s one can take $s_1/(s_2 - s_1^2/n)^{\frac{1}{2}}$, which is essentially Student's t-ratio. Theorem 7.1 implies then that in a sequential t-test the t-ratio at the nth stage is invariantly sufficient.

Example 7.2. (multiple correlation coefficient). Let $X_1, \cdots, X_n$ be independent and identically distributed according to a p-variate normal distribution ($p < n$), with unknown mean vector and unknown nonsingular covariance matrix. Let $X = (X_1, \cdots, X_n)$ so that X is a $p \times n$ matrix, and let Y be the matrix obtained from X by deleting the first row x of X. Define $\bar{X} = \sum X_\alpha/n$ (in this example α will always run from 1 to n) and $A = XX' - n\bar{X}\bar{X}'$, so that $\bar{X}$ and $A/(n-1)$ are the sample mean and sample covariance matrix. Similarly, define $\bar{Y} = \sum Y_\alpha/n$ and $B = YY' - n\bar{Y}\bar{Y}'$.

Suppose inference is desired about the population multiple correlation coefficient $\bar{R}^2$ between the first and the remaining variates. The corresponding sample multiple correlation coefficient R^2 can be written as [1] $R^2 = 1 -$

$|A|/(a_{11}|B|)$, in which a_{11} is the 11 element of A. Both $\bar{R}^2$ and R^2 are invariant under the group G composed of the following transformations:

(i) $X_\alpha \to X_\alpha + b$, b an arbitrary $p \times 1$ vector;

(ii) $Y \to CY$, C an arbitrary nonsingular $(p-1) \times (p-1)$ matrix;

(iii) $z \to cz$, c arbitrary real and $\neq 0$.

If the density of X is written down, it is seen to be of the form (7.1), with $\mathfrak{X} = pn$-space, and for s we may take the vector-valued function $(\bar{X}, A)$, whose components are the p components of $\bar{X}$ and the $\frac{1}{2}p(p+1)$ elements a_{ij} of A, with (say) $i \geq j$ (for simplicity of notation no distinction is made here between random variables and the values they may take on). All parts of Assumption C can be verified to hold. The only step that is not immediate is checking that the matrix $D(x)$ in (iii) of Assumption C is of maximal rank. It turns out that this condition is true on the set on which A is nonsingular, i.e. an $\mathfrak{A}_{SI}$ set of $\mathcal{P}$ measure 1. This can be shown by direct computation (facilitated by considering $\bar{X}$ and A functions of new variables, obtained from the $x_{i\alpha}$ by an orthogonal transformation), or by the following geometric argument: If the rows of X are projected on the $n-1$ dimensional orthogonal complement of the equiangular line, there results a matrix X^* such that $A = X^*X^{*\prime}$. The only nontrivial part of the proof is to show that the matrix of partial derivatives of the mapping $X^* \to A$ is of full rank. Now if the rows of X^* are linearly independent, then by the Gram Schmidt orthogonalization process we can write $X^* = TU$, where T is $p \times p$ lower triangular with positive diagonal elements, and U has orthonormal rows. We have then $A = TT'$, and the assertion follows from the fact that $X^* \to (T, U)$ and $T \to A$ are 1-1 bi-continuously differentiable.

Considering the transformations induced by G in the range of s, we can show readily that R^2 is a maximal invariant statistic based on s, so that it induces the σ-field $\mathfrak{A}_{SI}$. From the conclusion of Theorem 7.1 we know then that R^2 is invariantly sufficient. We can use this fact for a sequential test of a hypothesis concerning $\bar{R}^2$, by basing the test on the sequence of R^2 at the successive stages of sampling. The above stated result implies then that R^2 at the nth stage is invariantly sufficient.

The ordinary correlation coefficient between two variates can be treated in a completely analogous way.

EXAMPLE 7.3. Let $\mathfrak{X}$ be n space with the equiangular line deleted, and the function s as in Example 7.1. In contrast to the latter, let $p_\theta(x) = g_\theta(s(x))$, where $\{g_\theta\}$ consists of all positive measurable functions on $\mathcal{S}$ such that $\int g_\theta(s(x))\,dx = 1$. Denote $z = (1/\pi) \arctan |(s_2 - s_1^2/n)^{\frac{1}{2}}/s_1|$, so that $0 < z < 1$. Let G be the totality of transformations $x \to c(z)x$, where c is any positive analytic function on $[0, 1]$ such that $c(0) = 1$ (note that z is a function of x, and that the transformation does not change the value of x). Part (i) of Assumption C can be verified by direct computation; parts (ii) and (iii) are the same as in Example 7.1. The group G produces the same orbits as in Example 7.1, hence the same $\mathfrak{A}_I$. Therefore, the same conclusion obtains as in Example 7.1, i.e. the sequence of Student's ratios is an invariantly sufficient sequence.

In Example 7.3, Theorem 7.1 is easily applied. On the other hand, we cannot apply Theorem 3.1 directly, since we do not know how to verify Assumption A (ii) in this case due to the more complicated structure of G. Care has been taken in Example 7.3 that G does not contain an obvious subgroup that produces the same orbits and for which Assumption A (ii) can be verified (if there were such a subgroup, it could be used instead of G to yield the desired conclusion). The group of Example 7.1 is not a subgroup of G in Example 7.3 because G does not contain any transformation $x \to cx$ with c constant, except when $c = 1$. We can get a similar example by replacing the group (under multiplication) of functions $c(z)$ in Example 7.3 by the group of functions $c(z)$ defined by $\ln c(z) = \int (\exp zy)\alpha(dy)$, where α runs through the additive group of signed measures on the real line such that $\alpha(\{0\}) = 0$ and $\int (\exp y)|\alpha(dy)| < \infty$. The restriction $\alpha(\{0\}) = 0$ prevents the group of Example 7.1 from being a subgroup of G.

The essential difference between Assumptions A and C, as far as their verifiability is concerned, is that in Assumption A (ii) the structure of the group G comes into play, whereas in Assumption C conditions have to be verified only for each g separately. Example 7.3, even though admittedly artificial, shows that even when the family of distributions is dominated there may be cases where G is so complicated that the verification of Assumption A (ii) is either impossible or more difficult than the verification of Assumption C.

REFERENCES

[1] Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[2] Appleby, R. H. and Freund, R. J. (1962). An empirical evaluation of multivariate sequential procedure for testing means. *Ann. Math. Statist.* **33** 1413–1420.

[3] Armitage, P. (1957). Restricted sequential procedures. *Biometrika* **44** 9–26.

[4] Bahadur, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.* **25** 423–462.

[5] Barnard, G. A. (1952, 1953). The frequency justification of certain sequential tests. *Biometrika* **39** 144–150 and **40** 468–469.

[6] Benk, Robert H. (1964). Asymptotic properties of sequential probability ratio tests. Ph.D. thesis (unpublished), Harvard Univ.

[7] Burkholder, Donald L. (1960). The relation between sufficiency and invariance, I: theory. Invited address at the Central Regional Meeting of the Institute of Mathematical Statistics, Lafayette, Indiana.

[8] Cox, D. R. (1949). The use of the range in sequential analysis. *J. Roy. Statist. Soc. Ser. B* **11** 101–114.

[9] Cox, D. R. (1952). Sequential tests for composite hypotheses. *Proc. Cambridge Philos. Soc.* **48** 290–299.

[10] David, Herbert T. and Kruskal, William H. (1956). The WAGR sequential t-test reaches a decision with probability one. *Ann. Math. Statist.* **27** 797–805 and **29** 936.

[11] Doob, J. L. (1953). *Stochastic Processes*. Wiley, New York.

[12] Fraser, D. A. S. (1956). Sufficient statistics with nuisance parameters. *Ann. Math. Statist.* **27** 838–842.

[13] Fraser, D. A. S. (1957). *Nonparametric Methods in Statistics*. Wiley, New York.

[14] Ghosh, Bhaskar K. and Feeeman, Harold (1961). Introduction to Sequential Experimentation: Sequential Analysis of Variance. A monograph prepared by Groton

Associates, Groton, Massachusetts, for the Quartermaster Field Evaluation Agency, Fort Lee, Virginia.

[15] GHOSH, JAYANTA KUMAR (1960). On the monotonicity of the OC of a class of sequential probability ratio tests. *Calcutta Statist. Assoc. Bull.* **2** 139–144.

[16] GHOSH, JAYANTA KUMAR (1962). Optimum properties of sequential tests of simple and composite hypotheses and other related inference procedures. D. Phil. thesis (unpublished), Calcutta University.

[17] HAJNAL, J. (1961). A two-sample sequential *t*-test. *Biometrika* **48** 65–75.

[18] HALL, WM. JACKSON (1959). On sufficiency and invariance with applications in sequential analysis. Inst. of Statist. Mimeo. Ser. No. 228, Univ. of North Carolina, Chapel Hill. (contains errors).

[19] HALL, WM. JACKSON (1960). The relation between sufficiency and invariance, II: applications. Invited address at the Central Regional Meeting of the Institute of Mathematical Statistics, Lafayette, Indiana.

[20] HOEL, PAUL G. (1955). On a sequential test for the general linear hypothesis. *Ann. Math. Statist.* **26** 136–139.

[21] IPEAM, A. F. (1963). On the asymptotic behavior of densities with applications to several tests of composite hypotheses. Ph.D. thesis (unpublished), Univ. of Illinois.

[22] JACKSON, J. EDWARD and BRADLEY, RALPH A. (1961). Sequential χ^2- and T^2-tests. *Ann. Math. Statist.* **32** 1063–1077.

[23] JOHNSON, N. L. (1953). Some notes on the application of sequential methods in the analysis of variance. *Ann. Math. Statist.* **24** 614–623.

[24] JOHNSON, N. L. (1954). Sequential procedures in certain component of variance problems. *Ann. Math Statist.* **25** 357–366.

[25] JOHNSON, N. L. (1961). Sequential analysis: a survey. *J. Roy. Statist. Soc. Ser. A* **124** 372–411.

[26] KIEFER, J. (1957). Invariance, minimax sequential estimation, and continuous time processes. *Ann. Math. Statist.* **28** 573–601.

[27] KIEFER, J. and WEISS, LIONEL (1957). Some properties of generalized sequential probability ratio tests. *Ann. Math. Statist.* **28** 57–70, 14–17, and 72–74.

[28] LEHMANN, E. L. (1953). The power of rank tests. *Ann. Math. Statist.* **24** 23–43.

[29] LEHMANN, E. L. (1955). Ordered families of distributions. *Ann. Math. Statist.* **26** 399–419.

[30] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

[31] LEHMANN, E. L. and SCHEFFÉ, HENRY (1950). Completeness, similar regions, and unbiased estimation Part I. *Sankhyā* **10** 305–340.

[32] LOÈVE, MICHEL (1963). *Probability Theory* (3rd ed.). Van Nostrand, Princeton, N. J.

[33] NANDI, H. K. (1948). Use of well-known statistics in sequential analysis. *Sankhyā* **8** 339–344.

[34] NATIONAL BUREAU OF STANDARDS (1951). *Tables to Facilitate Sequential t-Tests*. Applied Math. Ser. 7, U. S. Government Printing Office, Washington.

[35] RAY, W. D. (1956). Sequential analysis applied to certain experimental designs in the analysis of variance. *Biometrika* **43** 388–403.

[36] RUSHTON, S. (1950). On a sequential *t*-test. *Biometrika* **37** 326–333.

[37] RUSHTON, S. (1952). On a two-sided sequential *t*-test. *Biometrika* **39** 302–308.

[38] RUSHTON, S. (1954). On the confluent hypergeometric function $M(\alpha, \gamma, x)$. *Sankhyā* **13** 369–376.

[39] RUSHTON, S. and LANG, E. D. (1954). Tables of the confluent hypergeometric function. *Sankhyā* **13** 377–411.

[40] SCHNEIDERMAN, M. A. and ARMITAGE, P. (1962). A family of closed sequential procedures. *Biometrika* **49** 41–56.

[41] SCHNEIDERMAN, M. A. and ARMITAGE, P. (1962). Closed sequential t-tests. *Biometrika* **49** 359–366.

[42] SLATER, L. J. (1960). *Confluent Hypergeometric Functions*. Cambridge Univ. Press.

[43] SOBEL, MILTON and WALD, ABRAHAM (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Ann. Math. Statist.* **20** 502–522.

[44] STEIN, CHARLES (1948). On sequences of experiments (abstract). *Ann. Math. Statist.* **19** 117–118.

[45] WALD, ABRAHAM (1947). *Sequential Analysis*. Wiley, New York.

[46] WILCOXON, FRANK, RHODES, L. J. and BRADLEY, RALPH A. (1963). Two sequential two-sample grouped rank tests with applications to screening experiments. *Biometrics* **19** 58–84.

[47] WIRJOSUDIRDJO, SUNARDI (1961). Limiting behavior of a sequence of density ratios. Ph.D. thesis (unpublished), Univ. of Illinois. Abstract in *Ann. Math. Statist.* **33** 296–297.