APPLICATION OF THE RADON-NIKODYM THEOREM TO THE THEORY OF SUFFICIENT STATISTICS[1]

By Paul R. Halmos[2] and L. J. Savage

University of Chicago

Summary. The body of this paper is written in terms of very general and abstract ideas which have been popular in pure mathematical work on the theory of probability for the last two or three decades. It seems to us that these ideas, so fruitful in pure mathematics, have something to contribute to mathematical statistics also, and this paper is an attempt to illustrate the sort of contribution we have in mind. The purpose of generality here is not to solve immediate practical problems, but rather to capture the logical essence of an important concept (sufficient statistic), and in particular to disentangle that concept from such ideas as Euclidean space, dimensionality, partial differentiation, and the distinction between continuous and discrete distributions, which seem to us extraneous.

In accordance with these principles the center of the stage is occupied by a completely abstract sample space—that is a set X of objects x, to be thought of as possible outcomes of an experimental program, distributed according to an unknown one of a certain set of probability measures. Perhaps the most familiar concrete example in statistics is the one in which X is n dimensional Cartesian space, the points of which represent n independent observations of a normally distributed random variable with unknown parameters, and in which the probability measures considered are those induced by the various common normal distributions of the individual observations.

A statistic is defined, as usual, to be a function T of the outcome, whose values, however, are not necessarily real numbers but may themselves be abstract entities. Thus, in the concrete example, the entire set of n observations, or, less trivially, the sequence of all sample moments about the origin are statistics with values in an n dimensional and in an infinite dimensional space respectively. Another illuminating and very general example of a statistic may be obtained as follows. Suppose that the outcomes of two not necessarily statistically independent programs are thought of as one united outcome—then the outcome T of the first program alone is a statistic relative to the united program. A technical measure theoretic result, known as the Radon-Nikodym theorem, is important in the study of statistics such as T. It is, for example, essential to the very definition of the basic concept of conditional probability of a subset E of X given a value y of T.

The statistic T is called sufficient for the given set $\mathfrak{M}$ of probability measures

---

[1] This paper was the basis of a lecture delivered upon invitation of the Institute at the meeting in Chicago on December 30, 1947.

[2] Fellow of the John Simon Guggenheim Memorial Foundation.

if (somewhat loosely speaking) the conditional probability of a subset E of X given a value y of T is the same for every probability measure in $\mathfrak{M}$. It is, for instance, well known that the sample mean and variance together form a sufficient statistic for the measures described in the concrete example.

The theory of sufficiency is in an especially satisfactory state for the case in which the set $\mathfrak{M}$ of probability measures satisfies a certain condition described by the technical term *dominated*. A set $\mathfrak{M}$ of probability measures is called dominated if each measure in the set may be expressed as the indefinite integral of a density function with respect to a fixed measure which is not itself necessarily in the set. It is easy to verify that both classical extremes, commonly referred to as the discrete and continuous cases, are dominated.

One possible formulation of the principal result concerning sufficiency for dominated sets is a direct generalization to the abstract case of the well known Fisher-Neyman result: T is sufficient if and only if the densities can be written as products of two factors, the first of which depends on the outcome through T only and the second of which is independent of the unknown measure. Another way of phrasing this result is to say that T is sufficient if and only if the likelihood ratio of every pair of measures in $\mathfrak{M}$ depends on the outcome through T only. The latter formulation makes sense even in the not necessarily dominated case but unfortunately it is not true in that case. The situation can be patched up somewhat by introducing a weaker notion called pairwise sufficiency.

In ordinary statistical parlance one often speaks of a statistic sufficient for some of several parameters. The abstract results mentioned above can undoubtedly be extended to treat this concept.

1. Basic definitions and notations. A measurable space $(X, \mathcal{S})$ is a set X and a σ-algebra $\mathcal{S}$ of subsets of X.[3] If $(X, \mathcal{S})$ and $(Y, \mathcal{T})$ are measurable spaces and if T is a transformation from X into Y (or, in other words, if T is a function with domain X and range in Y), then T is *measurable* if, for every F in $\mathcal{T}$, $T^{-1}(F) \in \mathcal{S}$. If Y is a Borel set in a finite dimensional Euclidean space, then we shall always understand that $\mathcal{T}$ is the class of all Borel subsets of Y, and the measurability of a function f from X to Y will be expressed by the notation $f(\epsilon)\ \mathcal{S}$.

Throughout most of what follows it will be assumed that $(X, \mathcal{S})$ and $(Y, \mathcal{T})$ are fixed measurable spaces and that T is a measurable transformation (also called a *statistic*) from X *onto* Y. A helpful example to keep in mind is the Cartesian plane in the role of X, its horizontal coordinate axis in the role of Y, and perpendicular projection from X onto Y in the role of T.

The following notations will be used. If g is a point function on Y (with arbitrary range), then gT is the point function on X defined by $gT(x) = g(T(x))$. If μ is a set function (with arbitrary range) on $\mathcal{S}$, then μT^{-1} is the set function

[3] A σ-algebra is a non empty class $\mathcal{S}$ of sets, closed under the formation of complements and countable unions. If $(X, \mathcal{S})$ is a measurable space, the sets of $\mathcal{S}$ will be called the measurable sets of X.

on $\boldsymbol{T}$ defined by $\mu T^{-1}(F) = \mu(T^{-1}(F))$. The class of all sets of the form $T^{-1}(F)$, with $F \in \boldsymbol{T}$, will be denoted by $T^{-1}(\boldsymbol{T})$; the characteristic function of a set A (in any space) will be denoted by χ_A.

LEMMA 1. *If g is any function on Y and A is any set in the range of g, then*

$$\{x: gT(x) \in A\} = T^{-1}(\{y: g(y) \in A\});$$

hence, in particular, $\chi_{T^{-1}(F)} = \chi_F T$ for every subset F of Y.[4]

PROOF. The following statements are mutually equivalent: (a) $x_0 \in \{x: gT(x) \in A\}$, (b) $g(T(x_0)) \in A$, (c) if $y_0 = T(x_0)$, then $g(y_0) \in A$, and (d) $T(x_0) \in \{y: g(y) \in A\}$. The equivalence of the first and last ones of these statements is exactly the assertion of the lemma.

We shall have frequent occasion to deal with functions on X which are induced by measurable functions on Y; the following result is a useful and direct structural characterization of such functions.

LEMMA 2. *If f is a real valued function on X, then a necessary and sufficient condition that there exist a measurable function g on Y such that $f = gT$ is that f $(\in)$ $T^{-1}(\boldsymbol{T})$; if such a function g exists, then it is unique.*[5]

PROOF. The necessity of the condition is clear. To prove sufficiency, suppose that f $(\in)$ $T^{-1}(\boldsymbol{T})$, $y_0 \in Y$, and write $X_0 = T^{-1}(\{y_0\})$. Suppose $x_0 \in X_0$ and write $E = \{x: f(x) = f(x_0)\}$. Since f $(\in)$ $T^{-1}(\boldsymbol{T})$, there is a set F in $\boldsymbol{T}$ such that $E = T^{-1}(F)$. Since $x_0 \in E$, it follows that $y_0 \in F$ and therefore that

$$X_0 = T^{-1}(\{y_0\}) \subset T^{-1}(F) = E.$$

In other words f is a constant on X_0 and consequently the equation $g(y_0) = f(x_0)$ unambiguously defines a function g on Y. The facts that $f = gT$ and that g is measurable are clear; the uniqueness of g follows from the fact that T maps X onto Y.

2. Measures and their derivatives. A *measure* is a real valued, non negative, finite (and therefore bounded), countably additive function on the measurable sets of a measurable space.[6] An integral whose domain of integration is not indicated is always to be extended over the whole space. If the symbol $[\mu]$, pronounced "modulo μ", follows an assertion concerning the points x of X, it is to be understood that the set E of those points for which the assertion is not true is such that $E \in \boldsymbol{S}$ and $\mu(E) = 0$. Thus, for instance, if f and g are functions (with arbitrary range) on X, then $f = g$ $[\mu]$ means that

[4] The symbol $\{- : -\}$ stands for the set of all those objects named before the colon which satisfy the condition stated after it.

[5] The notation f $(\in)$ $T^{-1}(\boldsymbol{T})$ means of course that f is a measurable function not only on the measurable space $(X, \boldsymbol{S})$ but also on the measurable space $(X, T^{-1}(\boldsymbol{T}))$. The restriction to real valued functions is inessential and is made only in order to avoid the introduction of more notation.

[6] Although most of the measures occurring in the applications of our theory are *probability measures* (i.e. measures whose value for the whole space is 1), the consideration of probability measures only is, in many of the proofs in the sequel, both unnecessary and insufficient.

$\mu(\{x: f(x) \neq g(x)\}) = 0$. Similarly, if f is a real valued function on X, then $f\ (\epsilon)\ T^{-1}(\boldsymbol{T})\ [\mu]$ means that there exists a real valued function g on X such that $g\ (\epsilon)\ T^{-1}(\boldsymbol{T})$ and $f = g\ [\mu]$.

If μ and ν are two measures on $\boldsymbol{S}$, ν is *absolutely continuous* with respect to μ, in symbols $\nu \ll \mu$, if $\nu(E) = 0$ for every measurable set E for which $\mu(E) = 0$. The measures μ and ν are *equivalent*, in symbols $\mu \equiv \nu$, if simultaneously $\mu \ll \nu$ and $\nu \ll \mu$.[7] One of the most useful results concerning absolute continuity is the Radon-Nikodym theorem, which may be stated as follows.[8]

A necessary and sufficient condition that $\nu \ll \mu$ is that there exist a non negative function f on X such that

$$\nu(E) = \int_E f(x)\ d\mu(x)$$

for every E in $\boldsymbol{S}$. The function f is unique in the sense that if also

$$\nu(E) = \int_E g(x)\ d\mu(x)$$

for every E in $\boldsymbol{S}$, then $f = g\ [\mu]$. If $\nu(E) \leqq \mu(E)$ for every E in $\boldsymbol{S}$, then $0 \leqq f(x) \leqq 1\ [\mu]$.

It is customary and suggestive to write $f = d\nu/d\mu$. Since $d\nu/d\mu$ is determined only to within a set for which μ vanishes, it follows that in a relation of the form

$$\frac{d\nu}{d\mu}\ (\epsilon)\ T^{-1}(\boldsymbol{T})\ [\mu]$$

the symbol $[\mu]$ is superfluous and may be omitted.

For typographical and heuristic reasons it is convenient sometimes to write the relation $f = d\nu/d\mu$ in the form $d\nu = f d\mu$; all the properties of Radon-Nikodym derivatives which are suggested by the well known differential formalism correspond to true theorems. Some of the ones that we shall make use of are trivial (e.g. $d\nu_1 = f_1 d\mu$ and $d\nu_2 = f_2 d\mu$ imply $d(\nu_1 + \nu_2) = (f_1 + f_2)d\mu$), while others are well known facts in integration theory (e.g. (i) $d\lambda = f d\nu$ and $d\nu = g d\mu$ imply $d\lambda = fg d\mu$, and (ii) $d\nu = f d\mu$ and $d\mu = g d\nu$ imply $fg = 1\ [\mu]$).

We conclude this section with a simple but useful result concerning the transformations of integrals.

LEMMA 3. *If g is a real valued function on Y and μ is a measure on $\boldsymbol{S}$, then*

$$\int_F g(y)\, d\mu T^{-1}(y) = \int_{T^{-1}(F)} gT(x)\ d\mu(x)$$

for every F in $\boldsymbol{T}$, in the sense that if either integral exists, then so does the other and the two are equal.

---

[7] It is clear that the relation of equivalence is reflexive, symmetric, and transitive, and hence deserves its name.

[8] For a proof of the Radon-Nikodym theorem and similar facts concerning the measure and integration theory which we employ, see S. Saks, *Theory of the Integral*, Warszawa—Lwów, 1937.

PROOF. Replacing g by $g\chi_F$ we see that it is sufficient to consider the case $F = Y$. The proof for this case follows from the observation that every approximating sum

$$\Sigma_i\, g(y_i)\mu T^{-1}(F_i)$$

of $\int g d\mu T^{-1}$ is also an approximating sum

$$\Sigma_i\, gT(x_i)\mu(E_i)$$

of $\int gT d\mu$, and conversely.[9]

3. Conditional probabilities and expectations. LEMMA 4. *If μ and ν are measures on $\mathbf{S}$ such that $\nu \ll \mu$, then $\nu T^{-1} \ll \mu T^{-1}$.*

PROOF. If $F \in \mathbf{T}$ and $0 = \mu T^{-1}(F) = \mu(T^{-1}(F))$, then

$$0 = \nu(T^{-1}(F)) = \nu T^{-1}(F).^{10}$$

Lemma 4 is the basis of the definition of a concept of great importance in probability theory. If μ is a measure on $\mathbf{S}$ and f is a non negative integrable function on X, then the measure ν defined by $d\nu = f d\mu$ is absolutely continuous with respect to μ. It follows from Lemma 4 that νT^{-1} is absolutely continuous with respect to μT^{-1}; we write $d\nu T^{-1} = g d\mu T^{-1}$. The function value $g(y)$ is known as the *conditional expectation* of f given y (or given that $T(x) = y$); we shall denote it by $e_\mu(f \mid y)$. If $f = \chi_E$ is the characteristic function of a set E in $\mathbf{S}$, then $e_\mu(f \mid y)$ is known as the *conditional probability* of E given y; we shall denote it by $p_\mu(E \mid y)$.[11]

The abstract nature of these definitions makes an intuitive justification of them desirable. Observe that since $\nu T^{-1}(F) = \nu(T^{-1}(F)) = \int_{T^{-1}(F)} f(x)\, d\mu(x)$, the defining equation of $e_\mu(f \mid y)$, written out in full detail, takes the form

$$\int_{T^{-1}(F)} f(x)\, d\mu(x) = \int_F e_\mu(f \mid y)\, d\mu T^{-1}(y), \qquad F \in \mathbf{T}.$$

---

[9] It is of interest to observe that either side of the equation in Lemma 3 may be obtained from the other by the formal substitution $y = T(x)$. A special case of this lemma is the celebrated and often misunderstood assertion that the expectation of a random variable is equal to the first moment of its distribution function.

[10] That the converse of Lemma 4 is not true is shown by the following example. Let X be the unit square, let Y be the unit interval, and let T be the perpendicular projection from X onto Y. Let μ be ordinary (Borel-Lebesgue) measure and let ν be linear measure on the intersection of X with, say, the horizontal line whose ordinate is $\frac{1}{2}$. Clearly ν is not absolutely continuous with respect to μ, but $\nu T^{-1} = \mu T^{-1}$.

[11] Definitions in this form were first proposed by A. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin, 1933. With a slight amount of additional trouble, conditional expectation could be defined for more general functions, but only the non negative case will occur in our applications.

If $f = \chi_E$, then this equation becomes the defining equation of $p_\mu(E \mid y)$:

$$\mu(E \cap T^{-1}(F)) = \int_F p_\mu(E \mid y)\, d\mu T^{-1}(y), \qquad F \in \mathcal{T}.^{12}$$

The customary definition of "the conditional probability of E given that $T(x) \in F$" is $\mu(E \cap T^{-1}(F))/\mu(T^{-1}(F))$, (assuming that the denominator does not vanish). Since $\mu(T^{-1}(F)) = \mu T^{-1}(F)$, we have

$$\frac{\mu(E \cap T^{-1}(F))}{\mu(T^{-1}(F))} = \frac{1}{\mu T^{-1}(F)} \int_F p_\mu(E \mid y)\, d\mu T^{-1}(y).$$

It is now formally plausible that if "F shrinks to a point y," then the left side of the last written equation should tend to the conditional probability of E given y and the right side should tend to the integrand $p_\mu(E \mid y)$. The use of the Radon-Nikodym differentiation theorem is a rigorous substitute for this rather shaky difference quotient approach.

Since $p_\mu(E \mid y)$ is determined, for each E, only to within a set for which μT^{-1} vanishes, it would be too optimistic to expect that, for each y, it behaves, regarded as a function of E, like a measure. It is, however, easy to prove that

(i) $p_\mu(X \mid y) = 1$ $[\mu T^{-1}]$,

(ii) $0 \leqq p_\mu(E \mid y) \leqq 1$ $[\mu T^{-1}]$,

(iii) if $\{E_n\}$ is a disjoint sequence of measurable sets, then $p_\mu(\bigcup_{n=1}^{\infty} E_n \mid y) = \sum_{n=1}^{\infty} p_\mu(E_n \mid y)$ $[\mu T^{-1}]$.[13]

The exceptional sets of measure zero depend in general on E in (ii) and on the particular sequence $\{E_n\}$ in (iii). It is interesting to observe that, despite the fact that μ need not be a probability measure, p_μ turns out always to have the normalization property (i). It is natural to ask whether or not the indeterminacy of $p_\mu(E \mid y)$ may be resolved, for each E, in such a way that the resulting function is a measure for each y, except possibly for a fixed set of y's on which μT^{-1} vanishes. Doob[14] has shown that this is the case when X is the real line; in the general case such a resolution is impossible. Fortunately, however, conditional probabilities are sufficiently tractable for most practical and theoretical purposes, and the requirement that they should behave like probability measures in the strict sense described above is almost never needed.

---

[12] We observe that it is not sufficient to require this for $F = Y$ only, i.e. to require $\mu(E) = \int p_\mu(E \mid y)\, d\mu T^{-1}(y)$. This special equation is satisfied by many functions which do not deserve the name conditional probability; e.g. it is satisfied by $p_\mu(E \mid y) = \text{constant} = \mu(E)/\mu T^{-1}(Y)$.

[13] See J. L. Doob, "Stochastic processes with an integral-valued parameter," *Am. Math. Soc. Trans.*, Vol. 44 (1938), pp. 95–98.

[14] See Doob, *loc. cit.* Doob asserts the theorem in much greater generality, but his proof is incorrect. The error in the proof and a counterexample to the general theorem were communicated to us by J. Dieudonné in a letter dated September 4, 1947. Doob's proof is valid for more general spaces than the real line (e.g. for finite dimensional Euclidean spaces and for compact metric spaces). The details of Dieudonné's counterexample will appear in a forthcoming book (entitled *Measure theory*) by Halmos.

We conclude this section with two easy but useful results which might also serve as illustrations of the method of finding conditional probabilities and expectations in certain special cases.

LEMMA 5. *If μ is a measure on $\mathbf{S}$, if g is a non negative function on Y, integrable with respect to μT^{-1}, and if ν is the measure on $\mathbf{S}$ defined by $d\nu = gTd\mu$, then $d\nu T^{-1} = gd\mu T^{-1}$, or, equivalently, $e_\mu(gT \mid y) = g(y)$ $[\mu T^{-1}]$.*

PROOF. From $\nu(E) = \int_E gT(x)\, d\mu(x)$ and Lemma 3 it follows that

$$\nu T^{-1}(F) = \nu(T^{-1}(F)) = \int_F g(y)\, d\mu T^{-1}(y).$$

LEMMA 6. *If μ is a measure on $\mathbf{S}$, if f and g are non negative functions on X and Y respectively, and if f, gT, and $f \cdot gT$ are all integrable with respect to μ, then*

$$e_\mu(f \cdot gT \mid y) = e_\mu(f \mid y) \cdot g(y) \quad [\mu T^{-1}].$$

Hence, in particular, if $F \in \mathbf{T}$, then

$$p_\mu(E \cap T^{-1}(F) \mid y) = p_\mu(E \mid y)\chi_F(y) \quad [\mu T^{-1}]$$

for every E in $\mathbf{S}$.

PROOF. If $d\nu = fd\mu$, then, by definition of e_μ, $\nu T^{-1}(F) = \int_F e_\mu(f \mid y)\, d\mu T^{-1}(y)$. Applications of Lemmas 3 and 5 yield

$$\int_F e_\mu(f \mid y)g(y)\, d\mu T^{-1}(y) = \int_F g(y)\, d\nu T^{-1}(y) = \int_{T^{-1}(F)} gT(x)\, d\nu(x)$$

$$= \int_{T^{-1}(F)} f(x)gT(x)\, d\mu(x) = \int_F e_\mu(f \cdot gT \mid y)\, d\mu T^{-1}(y),$$

and therefore the desired conclusion follows from the uniqueness assertion of the Radon-Nikodym theorem.

4. Dominated sets of measures. In many statistical situations it is necessary to consider simultaneously several measures on the same σ-algebra. The concept of absolute continuity is easily extended to sets of measures. If $\mathfrak{M}$ and $\mathfrak{N}$ are two sets of measures on $\mathbf{S}$ and if, for every set E in $\mathbf{S}$, the vanishing of $\mu(E)$ for every μ in $\mathfrak{M}$ implies the vanishing of $\nu(E)$ for every ν in $\mathfrak{N}$, then we shall call $\mathfrak{N}$ absolutely continuous with respect to $\mathfrak{M}$ and write $\mathfrak{N} \ll \mathfrak{M}$. If $\mathfrak{N} \ll \mathfrak{M}$ and $\mathfrak{M} \ll \mathfrak{N}$, the sets $\mathfrak{M}$ and $\mathfrak{N}$ are called equivalent and we write $\mathfrak{M} \equiv \mathfrak{N}$. If, in particular, $\mathfrak{M}$ contains exactly one measure μ, $\mathfrak{M} = \{\mu\}$, the abbreviated notations $\mathfrak{N} \ll \mu$, $\mu \ll \mathfrak{N}$, and $\mu \equiv \mathfrak{N}$, will be employed for $\mathfrak{N} \ll \mathfrak{M}$, $\mathfrak{M} \ll \mathfrak{N}$, and $\mathfrak{M} \equiv \mathfrak{N}$, respectively.

A set $\mathfrak{M}$ of measures on $\mathbf{S}$ will be called *dominated* if there exists a measure λ on $\mathbf{S}$ (not necessarily in $\mathfrak{M}$) such that $\mathfrak{M} \ll \lambda$. In applications there frequently occur sets of measures which are dominated in a sense apparently weaker than the one just defined—weaker in that the measure λ, which may for instance be

Lebesgue measure on the Borel sets of a finite dimensional Euclidean space, is not necessarily finite. It is easy to see, however, that whenever λ has the property (possessed by Lebesgue measure) that the space X is the union of countably many sets of finite measure, then a finite measure equivalent to λ exists and the two possible definitions of domination coincide.

The following result on dominated sets of measures may be found to have some interest of its own and will be applied in the sequel.

LEMMA 7. *Every dominated set of measures has an equivalent countable subset.*

PROOF. Let $\mathfrak{M}$ be a dominated set of measures on S, $\mathfrak{M} \ll \lambda$; for any μ in $\mathfrak{M}$ write $f_\mu = d\mu/d\lambda$ and $K_\mu = \{x: f_\mu(x) > 0\}$. We define (for the purposes of this proof only) a *kernel* as a set K in S such that, for some measure μ in $\mathfrak{M}$, $K \subset K_\mu$ and $\mu(K) > 0$; we define a *chain* as a disjoint union of kernels. Since $\lambda(K) > 0$ for every kernel K, it follows from the finiteness of λ that every chain is a *countable* disjoint union of kernels. It follows also from these definitions that if C is a measurable subset of a chain, such that $\mu(C) > 0$ for at least one measure μ in $\mathfrak{M}$, then C is a chain, and that a disjoint union of chains is a chain. The last two remarks imply, through the usual process of disjointing any countable union, that a countable (but not necessarily disjoint) union of chains is a chain.

Let $\{C_j\}$ be a sequence of chains such that, as $j \to \infty$, $\lambda(C_j)$ approaches the supremum of the values of λ on chains. If $C = \bigcup_{i=1}^{\infty} C_j$, then C is a chain for which $\lambda(C)$ is maximal. The definition of a chain yields the existence of a sequence $\{K_i\}$ of kernels such that $C = \bigcup_{i=1}^{\infty} K_i$, and the definition of a kernel yields the existence, for each $i = 1, 2, \cdots$, of a measure μ_i in $\mathfrak{M}$ such that $K_i \subset K_{\mu_i}$ and $\mu_i(K_i) > 0$. We write $\mathfrak{N} = \{\mu_1, \mu_2, \cdots\}$; since $\mathfrak{N} \subset \mathfrak{M}$, the relation $\mathfrak{N} \ll \mathfrak{M}$ is trivial. We shall prove that $\mathfrak{M} \ll \mathfrak{N}$.

Suppose that $E \in S$, $\mu_i(E) = 0$ for $i = 1, 2, \cdots$, and let μ be any measure in $\mathfrak{M}$. It is to be proved that $\mu(E) = 0$. Since $\mu(E - K_\mu) = 0$, there is no loss of generality in assuming that $E \subset K_\mu$. If $\mu(E - C) > 0$, then $\lambda(E - C) > 0$ and therefore (since $E - C$ is a kernel) $E \cup C$ is a chain with $\lambda(E \cup C) > \lambda(C)$. Since this is impossible, it follows that $\mu(E - C) = 0$. Since $0 = \mu_i(E) = \mu_i(E \cap K_i) = \int_{E \cap K_i} f_{\mu_i}\, d\lambda$ and since $K_i \subset K_{\mu_i}$, it follows that $\lambda(E \cap K_i) = 0$. We conclude that $\lambda(E \cap C) = \sum_{i=1}^{\infty} \lambda(E \cap K_i) = 0$ and therefore $\mu(E \cap C) = 0$. Since $\mu(E) = \mu(E - C) + \mu(E \cap C)$, the proof of the lemma is complete.

5. Sufficient statistics for dominated sets. The statistic T is sufficient for a set $\mathfrak{M}$ of measures on S if, for every E in S, there exists a measurable function $p = p(E \mid y)$ on Y, such that

$$p_\mu(E \mid y) = p(E \mid y) \; [\mu T^{-1}]$$

for every μ in $\mathfrak{M}$.[15] In other words, T is sufficient for $\mathfrak{M}$ if there exists a condi-

[15] The original definition of sufficiency was given by R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Roy. Soc. Phil. Trans.*, Series A, Vol. 222 (1922), pp. 309–368.

tional probability function common to every μ in $\mathfrak{M}$, or, crudely speaking, if the conditional distribution induced by T is independent of μ.

THEOREM 1. *A necessary and sufficient condition that the statistic T be sufficient for a dominated set $\mathfrak{M}$ of measures on $\mathbf{S}$ is that there exist a measure λ on $\mathbf{S}$ such that $\mathfrak{M} \equiv \lambda$ and such that $d\mu/d\lambda$ (ϵ) $T^{-1}(\mathbf{T})$ for every μ in $\mathfrak{M}$.*

Proof of necessity. Let $\mathfrak{N} = \{\mu_1, \mu_2, \cdots\}$ be a countable subset equivalent to $\mathfrak{M}$ (Lemma 7), and write λ for the measure on $\mathbf{S}$ defined by

$$\lambda(E) = \sum_{i=1}^{\infty} a_i \mu_i(E),$$

where $a_i = 1/2^i \mu_i(X)$, $i = 1, 2, \cdots$. Clearly $\mathfrak{M} \equiv \lambda$.

If p is a conditional probability function common to every μ in $\mathfrak{M}$, then, for every F in $\mathbf{T}$,

$$\lambda(E \cap T^{-1}(F)) = \sum_{i=1}^{\infty} a_i \mu_i(E \cap T^{-1}(F))$$

$$= \sum_{i=1}^{\infty} a_i \int_F p(E \mid y)\, d\mu_i T^{-1}(y) = \int_F p(E \mid y)\, d\lambda T^{-1}(y),$$

i.e. p serves also as a conditional probability for λ.

Take any fixed μ in $\mathfrak{M}$, write $d\mu/d\lambda = f$, and $e_\lambda(f \mid y) = g(y)$; then $d\mu T^{-1} = g d\lambda T^{-1}$, and we have, for every E in $\mathbf{S}$,

$$\int_E f(x)\, d\lambda(x) = \mu(E) = \int p(E \mid y)\, d\mu T^{-1}(y)$$

$$= \int p(E \mid y) g(y)\, d\lambda T^{-1}(y) = \int e_\lambda(\chi_E \mid y) e_\lambda(gT \mid y)\, d\lambda T^{-1}(y)$$

$$= \int e_\lambda(\chi_E \cdot gT \mid y)\, d\lambda T^{-1}(y) = \int \chi_E(x) gT(x)\, d\lambda(x) = \int_E gT(x)\, d\lambda(x).$$

The desired result, $f(x) = gT(x)$ $[\lambda]$, follows from a comparison of the first and last terms in the last written chain of equations.

Proof of Sufficiency. We shall prove that p_λ is a conditional probability function common to every μ in $\mathfrak{M}$. Take any fixed E in $\mathbf{S}$ and μ in $\mathfrak{M}$ and write $d\mu/d\lambda = gT$. If the measure ν is defined by $d\nu = \chi_E d\mu$, then $d\nu T^{-1} = p_\mu d\mu T^{-1}$, where $p_\mu = p_\mu(E \mid y)$. The hypothesis $d\mu = gT d\lambda$ implies that $d\mu T^{-1} = g d\lambda T^{-1}$ and hence that

$$d\nu T^{-1} = p_\mu \cdot g d\lambda T^{-1}.$$

On the other hand $d\nu = \chi_E d\mu = \chi_E \cdot gT d\lambda$, so that

$$d\nu T^{-1} = e_\lambda d\lambda T^{-1},$$

where $e_\lambda = e_\lambda(\chi_E \cdot gT \mid y) = p_\lambda(E \mid y) g(y)$. It follows from a comparison of the two expressions for $d\nu T^{-1}$ that

$$p_\mu(E \mid y) g(y) = p_\lambda(E \mid y) g(y)\ [\lambda T^{-1}].$$

Since the relation $d\mu T^{-1} = gd\lambda T^{-1}$ clearly implies that $g(y) \neq 0$ $[\mu T^{-1}]$ (i.e. that $\mu T^{-1}(\{y: g(y) = 0\}) = 0$), it follows, finally that

$$p_\mu(E \mid y) = p_\lambda(E \mid y) \; [\mu T^{-1}].$$

6. Special criteria for sufficiency. Theorem 1 may be recast in a form more akin in spirit to previous investigations of the concept of sufficiency.[16]

COROLLARY 1. *A necessary and sufficient condition that the statistic T be sufficient for a dominated set $\mathfrak{M}$ ($\ll \lambda_0$) of measures on $\mathbf{S}$ is that, for every μ in $\mathfrak{M}$, $f_\mu = d\mu/d\lambda_0$ be factorable in the form $f_\mu = g_\mu \cdot t$, where $0 \leqq g_\mu$ (ϵ) $T^{-1}(\mathbf{T})$, $0 \leqq t$, t and $g_\mu \cdot t$ are integrable with respect to λ_0, and t vanishes $[\lambda_0]$ on each set in $\mathbf{S}$ for which every μ in $\mathfrak{M}$ vanishes.*

In more customary statistical language the condition asserts essentially that "each density is factorable into a function of the statistic alone and a function independent of the parameter."

PROOF. If T is sufficient for $\mathfrak{M}$, then there exists a measure λ with the properties described in Theorem 1. It follows that

$$f_\mu = \frac{d\mu}{d\lambda_0} = \frac{d\mu}{d\lambda} \frac{d\lambda}{d\lambda_0}$$

and we may write $g_\mu = d\mu/d\lambda$ and $t = d\lambda/d\lambda_0$. The only assertion that is not immediately obvious is the one concerning the vanishing of t. To prove it, suppose that $\mu(E) = 0$ for every μ in $\mathfrak{M}$; the fact that then

$$0 = \lambda(E) = \int_E t(x) \, d\lambda_0(x)$$

implies the desired conclusion.

If, conversely, $f_\mu = g_\mu \cdot t$, then we may write $d\lambda = td\lambda_0$. The relation $\mathfrak{M} \equiv \lambda$ follows from the statement concerning the vanishing of t, and the relation $d\mu/d\lambda$ (ϵ) $T^{-1}(\mathbf{T})$ is implied by the equation $d\mu = g_\mu \cdot td\lambda_0 = g_\mu d\lambda$.

For the statement of the next consequence of Theorem 1 it is convenient to call a set $\mathfrak{M}$ of measures on $\mathbf{S}$ *homogeneous* if $\mu \equiv \nu$ for every μ and ν in $\mathfrak{M}$.

COROLLARY 2. *A necessary and sufficient condition that the statistic T be sufficient for a homogeneous set $\mathfrak{M}$ of measures on $\mathbf{S}$ is that, for every μ and ν in $\mathfrak{M}$, $d\nu/d\mu$ (ϵ) $T^{-1}(\mathbf{T})$.*

PROOF. Since a homogeneous set is dominated (by any one of its elements), Theorem 1 is applicable. If T is sufficient for $\mathfrak{M}$ and if λ has the properties described in Theorem 1, then $d\nu/d\mu = (d\nu/d\lambda)/(d\mu/d\lambda)$. The converse follows, through Theorem 1, by letting λ be any measure in $\mathfrak{M}$.

We shall say that the statistic T is *pairwise sufficient* for a set $\mathfrak{M}$ of measures

[16] See J. Neyman, "Su un teorema concernente le cosiddette statistiche sufficienti," *Inst. Ital. Atti. Giorn.*, Vol. 6 (1935), pp. 320–334. In this paper Neyman is somewhat restricted by his use of classical analytical methods, but he points out the possibility and desirability of extending his results to a much more general domain. For a recent presentation of the theory and further references to the literature cf. H. Cramér, *Mathematical Methods of Statistics*, Princeton, 1946.

on $\boldsymbol{S}$ if it is sufficient for every pair $\{\mu, \nu\}$ of measures in $\mathfrak{M}$. In other words, T is pairwise sufficient for $\mathfrak{M}$ if, for every E in $\boldsymbol{S}$ and μ and ν in $\mathfrak{M}$, there exists a measurable function $p_{\mu\nu}(E \mid y)$ on Y such that

$$p_\mu(E \mid y) = p_{\mu\nu}(E \mid y) \ [\mu T^{-1}] \qquad \text{and} \qquad p_\nu(E \mid y) = p_{\mu\nu}(E \mid y) \ [\nu T^{-1}].$$

Since pairwise sufficiency is (at least apparently) weaker than sufficiency, it is not surprising that there is a simple criterion for it even in the case of quite arbitrary (not necessarily homogeneous or dominated) sets of measures.

COROLLARY 3. *A necessary and sufficient condition that T be pairwise sufficient for a set $\mathfrak{M}$ of measures on $\boldsymbol{S}$ is that, for any two measures μ and ν in $\mathfrak{M}$, $d\mu/d(\mu + \nu)$ (ϵ) $T^{-1}(\boldsymbol{T})$.*

PROOF. If T is sufficient for μ and ν, then there exists a measure $\lambda \equiv \mu + \nu$ such that $d\mu/d\lambda$ (ϵ) $T^{-1}(\boldsymbol{T})$ and $d\nu/d\lambda$ (ϵ) $T^{-1}(\boldsymbol{T})$. It follows that

$$\frac{d\mu}{d(\mu + \nu)} = \frac{d\mu}{d\lambda} \Big/ \frac{d(\mu + \nu)}{d\lambda} = \frac{d\mu}{d\lambda} \Big/ \left(\frac{d\mu}{d\lambda} + \frac{d\nu}{d\lambda}\right).$$

The sufficiency of the condition follows immediately by applying Theorem 1 to the two-element set $\{\mu, \nu\}$.

7. Pairwise sufficiency and likelihood ratios. It is sometimes convenient to express the result of Corollary 3 in slightly different language. If λ is a measure on $\boldsymbol{S}$ and if f and g are real valued measurable functions on X such that $\lambda(\{x: f(x) = g(x) = 0\}) = 0$, we shall say that the pair (f, g) is *admissible* $[\lambda]$. (Intuitively an admissible pair (f, g) is to be thought of as a ratio f/g, which, however, may not be formed directly at the points x for which $g(x) = 0$.) Two admissible pairs (f_1, g_1) and (f_2, g_2) will be called *equivalent* $[\lambda]$, in symbols $(f_1, g_1) \equiv (f_2, g_2)$ $[\lambda]$, if there exists a real valued measurable function t on X such that $t(x) \neq 0$ $[\lambda]$ and such that $f_1 = tf_2$ and $g_1 = tg_2$ $[\lambda]$. It is clear that the relation " $\equiv$ $[\lambda]$" is indeed an equivalence; the equivalence class containing the admissible pair (f, g) will be called the *ratio* of f and g and will be denoted by $f \mid g$. (A ratio may accordingly be described as a measurable function from X to the real projective line.) For a ratio $f \mid g$ we shall write $f \mid g$ (ϵ) $T^{-1}(\boldsymbol{T})$ $[\lambda]$ if the equivalence class $f \mid g$ contains a pair (f_0, g_0) which is admissible $[\lambda]$ and for which f_0 (ϵ) $T^{-1}(\boldsymbol{T})$ and g_0 (ϵ) $T^{-1}(\boldsymbol{T})$.

LEMMA 8. *If μ, ν, λ_1, and λ_2 are measures on $\boldsymbol{S}$ such that $\mu + \nu \ll \lambda_1$ and $\mu + \nu \ll \lambda_2$, then the pairs $(d\mu/d\lambda_1, d\nu/d\lambda_1)$ and $(d\mu/d\lambda_2, d\nu/d\lambda_2)$ are admissible $[\mu + \nu]$ and equivalent $[\mu + \nu]$.*

PROOF. The admissibility of, for instance, $(d\mu/d\lambda_1, d\nu/d\lambda_1)$ follows from the fact that $d\mu/d\lambda_1 \neq 0$ $[\mu]$ and $d\nu/d\lambda_1 \neq 0$ $[\nu]$, whence

$$(\mu + \nu)\left(\left\{x: \frac{d\mu}{d\lambda_1}(x) = \frac{d\nu}{d\lambda_1}(x) = 0\right\}\right) = 0.$$

To prove equivalence, we write $\lambda_1 + \lambda_2 = \lambda$. Since

$$\frac{d\mu}{d\lambda_1}\frac{d\lambda_1}{d\lambda} = \frac{d\mu}{d\lambda} = \frac{d\mu}{d\lambda_2}\frac{d\lambda_2}{d\lambda}, \qquad \frac{d\nu}{d\lambda_1}\frac{d\lambda_1}{d\lambda} = \frac{d\nu}{d\lambda} = \frac{d\nu}{d\lambda_2}\frac{d\lambda_2}{d\lambda},$$

since also $d\lambda_1/d\lambda \neq 0$ $[\lambda_1]$ and therefore $d\lambda_1/d\lambda \neq 0$ $[\mu + \nu]$, and since, similarly, $d\lambda_2/d\lambda \neq 0$ $[\mu + \nu]$, the conditions of the definition of equivalence are satisfied by $t = (d\lambda_2/d\lambda)/(d\lambda_1/d\lambda)$.

If μ and ν are any two measures on $\mathbf{S}$ and if λ is any measure on $\mathbf{S}$ such that $\mu + \nu \ll \lambda$ (for instance if $\lambda = \mu + \nu$), then the ratio $d\mu/d\lambda \mid d\nu/d\lambda$, which according to Lemma 8 exists $[\mu + \nu]$ and is independent of λ, will be called the *likelihood ratio* of μ and ν and will be denoted by $d\mu \mid d\nu$. The result of Corollary 3 may be expressed in terms of likelihood ratios as follows.

THEOREM 2. *A necessary and sufficient condition that T be pairwise sufficient for a set $\mathfrak{M}$ of measures on $\mathbf{S}$ is that, for any two measures μ and ν in $\mathfrak{M}$, $d\mu \mid d\nu$ (ϵ) $T^{-1}(\boldsymbol{T})$.*

PROOF. If T is sufficient for μ and ν, then, by Corollary 3, $d\mu/d(\mu + \nu)$ (ϵ) $T^{-1}(\boldsymbol{T})$, $d\nu/d(\mu + \nu)$ (ϵ) $T^{-1}(\boldsymbol{T})$, and, by Lemma 8, $(d\mu/d(\mu + \nu), d\nu/d(\mu + \nu))$ is an admissible pair belonging to the equivalence class $d\mu \mid d\nu$. Suppose conversely that $f = d\mu/d(\mu + \nu)$, $g = d\nu/d(\mu + \nu)$, and let the real valued measurable functions t, f_0, and g_0 be such that $t \neq 0$ $[\mu + \nu]$, f_0 (ϵ) $T^{-1}(\boldsymbol{T})$, g_0 (ϵ) $T^{-1}(\boldsymbol{T})$, (f_0, g_0) is admissible $[\mu + \nu]$, and

$$f = t \cdot f_0, \qquad g = t \cdot g_0 \; [\mu + \nu].$$

Since f and g are non negative, it follows that $f = |t| \cdot |f_0|$ and $g = |t| \cdot |g_0|$ $[\mu + \nu]$, i.e. that there is no loss of generality in assuming that t, f_0, and g_0 are non negative. The relation $f + g = 1$ $[\mu + \nu]$ implies that $t \cdot (f_0 + g_0) = 1$ $[\mu + \nu]$; the fact that (f_0, g_0) is admissible $[\mu + \nu]$ then yields $t \; \epsilon \; T^{-1}(\boldsymbol{T})$. The proof is completed by comparing this result with the expressions for f and g in terms of f_θ and g_0 and applying Corollary 3.

8. Pairwise sufficiency versus sufficiency. In order to show that our results on pairwise sufficiency (in the preceding section and in the sequel) are not vacuous, we proceed now to exhibit a statistic which is, for a suitable set of measures, pairwise sufficient but not sufficient.

Let $X = \{(x, i): 0 \leqq x \leqq 1, i = 0, 1\}$ be the union of two unit intervals and let $Y = \{y: 0 \leqq y \leqq 1\}$ be a unit interval. In accordance with our basic convention, measurability in both X and Y is to be taken in the sense of Borel. The statistic T is defined by $T(x, i) = x$.

Write $X_0 = \{(x, 0): 0 \leqq x \leqq 1\}$ and $X_1 = \{(x, 1): 0 \leqq x \leqq 1\}$. Let μ be (linear) Lebesgue measure on the class $\mathbf{S}$ of Borel subsets of X, and define, whenever $E \; \epsilon \; \mathbf{S}$ and $0 \leqq \alpha \leqq 1$,

$$\mu_\alpha(E) = \tfrac{1}{2}[\mu(E \cap X_0) + \chi_{E \cap X_1}(\alpha, 1)].$$

Let ν be (linear) Lebesgue measure on the class $\boldsymbol{T}$ of Borel subsets of Y, and define, whenever $F \; \epsilon \; \mathbf{T}$ and $0 \leqq \alpha \leqq 1$,

$$\nu_\alpha(F) = \tfrac{1}{2}[\nu(F) + \chi_F(\alpha)].$$

Clearly $\nu_\alpha = \mu_\alpha T^{-1}$; we write $\mathfrak{M} = \{\mu_\alpha : 0 \leqq \alpha \leqq 1\}$.

If $\delta(y, \alpha)$ is defined to be 1 or 0 according as $y = \alpha$ or $y \neq \alpha$, if $\delta'(y, \alpha) = 1 - \delta(y, \alpha)$, and if

$$p_\alpha(E \mid y) = \delta'(y, \alpha)\chi_E(y, 0) + \delta(y, \alpha)\chi_E(y, 1),$$

then a straightforward computation shows that

$$\mu_\alpha(E \cap T^{-1}(F)) = \int_F p_\alpha(E \mid y)\, d\nu_\alpha(y),$$

so that $p_\alpha(E \mid y) = p_{\mu_\alpha}(E \mid y)$ $[\nu_\alpha]$.

It is now easy to verify that T is pairwise sufficient for $\mathfrak{M}$. Indeed if α and β are any two different numbers in the closed unit interval, we may write

$$p(E \mid y) = \delta'(y, \alpha)\delta'(y, \beta)\chi_E(y, 0) + [\delta(y, \alpha) + \delta(y, \beta)]\chi_E(y, 1).$$

Since $\{y: p(E \mid y) \neq p_\alpha(E \mid y)\} = \{\beta\}$ and $\{y: p(E \mid y) \neq p_\beta(E \mid y)\} = \{\alpha\}$, it follows that $p(E \mid y) = p_\alpha(E \mid y)$ $[\nu_\alpha]$ and $p(E \mid y) = p_\beta(E \mid y)$ $[\nu_\beta]$.

To prove that T is not sufficient for $\mathfrak{M}$ we observe that $p_\alpha(X_1 \mid y) = \delta(y, \alpha)\chi_{X_1}(y, 1) = \delta(y, \alpha)$ and therefore

$$p_{\mu_\alpha}(X_1 \mid y) = \delta(y, \alpha)\ [\nu_\alpha].$$

Suppose that there is a conditional probability function p such that $p(E \mid y) = p_{\mu_\alpha}(E \mid y)$ $[\nu_\alpha]$. Then, in particular,

$$p(X_1 \mid y) = \delta(y, \alpha)\ [\nu_\alpha].$$

Since $\nu_\alpha(\{\alpha\}) = \frac{1}{2} > 0$, it follows that

$$p(X_1 \mid \alpha) = \delta(\alpha, \alpha) = 1,$$

or, changing to a more suggestive notation, that $p(X_1 \mid y) = 1$ for all y. We have, however,

$$\begin{aligned}\nu_\alpha(\{y: p_\alpha(X_1 \mid y) = 0\}) &= \nu_\alpha(\{y: \delta(y, \alpha) = 0\}) \\ &= \nu_\alpha(\{y: y \neq \alpha\}) = \tfrac{1}{2},\end{aligned}$$

so that $\nu_\alpha(\{y: p_{\mu_\alpha}(X_1 \mid y) = 0\}) = \frac{1}{2}$. This contradiction shows the impossibility of the existence of a conditional probability function common to every μ in $\mathfrak{M}$.

This example shows also that, in a sense, sufficiency is more fundamental than pairwise sufficiency. If, for instance, we imagine that it is important to a statistician that he either estimate α sharply or refrain from estimating it altogether, then he is by no means as well off with the observation of y as with that of x.

9. Pairwise sufficiency for dominated sets. We now proceed to show that for dominated sets of measures no such example as the one in the preceding section exists, or, in other words, that for dominated sets the concepts of pairwise sufficiency and sufficiency do coincide.

LEMMA 9. *If T is pairwise sufficient for a set $\{\mu_0, \mu_1, \mu_2\}$ of three measures on $\mathbf{S}$, then*[17]

$$\frac{d\mu_0}{d(\mu_0 + \mu_1 + \mu_2)} (\epsilon)T^{-1}(\boldsymbol{T}).$$

PROOF. According to Corollary 3,

$$f_1 = \frac{d\mu_0}{d(\mu_0 + \mu_1)} (\epsilon)T^{-1}(\boldsymbol{T}) \quad \text{and} \quad f_2 = \frac{d\mu_0}{d(\mu_0 + \mu_2)} (\epsilon)T^{-1}(\boldsymbol{T}).$$

Since $d\mu_0 = f_1 d(\mu_0 + \mu_1) = f_2 d(\mu_0 + \mu_2)$, we have $f_1 d\mu_0 = f_1 f_2 d(\mu_0 + \mu_2)$ and $f_2 d\mu_0 = f_1 f_2 d(\mu_0 + \mu_1)$, so that

$$(f_1 + f_2 - f_1 f_2) d\mu_0 = f_1 f_2 d(\mu_0 + \mu_1 + \mu_2).$$

If we write $d\mu_0 = f d(\mu_0 + \mu_1 + \mu_2)$, then it follows that

$$(f_1 + f_2 - f_1 f_2) f = f_1 f_2 \, [\mu_0 + \mu_1 + \mu_2].$$

Since $0 \leqq f_1 \leqq 1$ and $0 \leqq f_2 \leqq 1$, the equation $f_1 + f_2 - f_1 f_2 = 0$ is equivalent to $f_1 = f_2 = 0$. Since $\mu_0(\{x: f_1(x) = f_2(x) = 0\}) = 0$, it follows that f may be redefined, if necessary, to be 0 on the set $\{x: f_1(x) = f_2(x) = 0\}$ without affecting the relation $d\mu_0 = f d(\mu_0 + \mu_1 + \mu_2)$; since outside this set $f = f_1 f_2 / (f_1 + f_2 - f_1 f_2)$, the proof of the lemma is complete.

LEMMA 10. *If T is pairwise sufficient for a finite set $\{\mu_0, \mu_1, \cdots, \mu_k\}$ of measures on $\mathbf{S}$, then $d\mu_0 / d(\sum_{i=0}^{k} \mu_i)$ (ϵ) $T^{-1}(\boldsymbol{T})$.*

PROOF. For $k = 1$ the conclusion is a restatement of the hypothesis; we proceed by induction. Given $\mu_0, \mu_1, \cdots, \mu_{k+1}$, we write $\mu = \sum_{i=1}^{k} \mu_i$. Then $d\mu_0 / d(\mu_0 + \mu)$ (ϵ) $T^{-1}(\boldsymbol{T})$ by the induction hypothesis and $d\mu_0 / d(\mu_0 + \mu_{k+1})$ (ϵ) $T^{-1}(\boldsymbol{T})$ by Corollary 3. Lemma 9 may then be applied to $\{\mu_0, \mu, \mu_{k+1}\}$ and yields the desired conclusion.

LEMMA 11. *If $\{\mu_0, \mu_1, \mu_2 \cdots\}$ is a sequence of measures on $\mathbf{S}$ such that $\sum_{i=0}^{\infty} \mu_i(X) < \infty$; if, for every E in $\mathbf{S}$, $\mu(E) = \sum_{i=0}^{\infty} \mu_i(E)$; and if λ is a measure $\mathbf{S}$ such that $\mu_i \ll \lambda$ for $i = 0, 1, 2, \cdots$, then*

$$\lim_k d(\textstyle\sum_{i=0}^{k} \mu_i)/d\lambda = d\mu/d\lambda \; [\lambda].$$

PROOF. Since $0 \leqq d(\sum_{i=0}^{k} \mu_i)/d\lambda = \sum_{i=0}^{k} (d\mu_i/d\lambda) \leq d\mu/d\lambda \; [\lambda]$, the series $\sum_{i=0}^{\infty} (d\mu_i/d\lambda)$ does indeed converge to a measurable function f $[\lambda]$. Since, for every E in $\mathbf{S}$,

$$\int_E f \, d\lambda = \sum_{i=0}^{\infty} \int_E \frac{d\mu_i}{d\lambda} \, d\lambda = \sum_{i=0}^{\infty} \mu_i(E) = \mu(E),$$

we have $f = d\mu/d\lambda$ $[\lambda]$, as stated.

[17] In view of Theorem 1, Lemma 9 asserts that if T is pairwise sufficient for a set $\mathfrak{M}$ of three elements, then T is sufficient for $\mathfrak{M}$. Lemmas 10 and 12 extend this result to finite and countably infinite sets $\mathfrak{M}$ respectively. Since every countable set of measures is dominated, the final result, Theorem 3, contains all these preliminaries as special cases.

LEMMA 12. *If* $\{\mu_0, \mu_1, \mu_2, \cdots\}$ *is a sequence of measures on* $\mathbf{S}$ *such that* $\sum_{i=0}^{\infty} \mu_i(X) < \infty$, *and if, for every* E *in* $\mathbf{S}$, $\mu(E) = \sum_{i=0}^{\infty} \mu_i(E)$, *then*

$$\lim_k d\mu_0/d(\textstyle\sum_{i=0}^{k} \mu_i) = d\mu_0/d\mu \ [\mu].$$

If, in addition, T *is pairwise sufficient for the sequence* $\{\mu_0, \mu_1, \mu_2 \cdots\}$, *then* $d\mu_0/d\mu$ (ϵ) $T^{-1}(\boldsymbol{T})$.

PROOF. We have, for $k = 0, 1, 2, \cdots$,

$$\frac{d\mu_0}{d(\sum_{i=0}^{k} \mu_i)} \cdot \frac{d(\sum_{i=0}^{k} \mu_i)}{d\mu} = \frac{d\mu_0}{d\mu}.$$

If we write $\lambda = \mu$, then the hypotheses of Lemma 11 are satisfied and, consequently, the second factor on the left side converges to 1 $[\mu]$; it follows that the first factor converges to $d\mu_0/d\mu$ $[\mu]$. The second assertion of the lemma follows from Lemma 10.

THEOREM 3. *A necessary and sufficient condition that* T *be sufficient for a dominated set* $\mathfrak{M}$ *of measures on* $\mathbf{S}$ *is that* T *be pairwise sufficient for* $\mathfrak{M}$.

PROOF. The necessity of the condition is obvious. To prove its sufficiency, let $\mathfrak{N} = \{\mu_1, \mu_2, \cdots\}$ be a countable subset of $\mathfrak{M}$ which is equivalent to $\mathfrak{M}$ (Lemma 7), and let μ_0 be an arbitrary measure in $\mathfrak{M}$. Since the sufficiency or pairwise sufficiency of T remains unaltered if some or all of the measures in $\mathfrak{M}$ are replaced by positive constant multiples of themselves, we may assume that $\sum_{i=0}^{\infty} \mu_i(X) < \infty$. If we write, for every E in $\mathbf{S}$, $\lambda(E) = \sum_{i=1}^{\infty} \mu_i(E)$, then the pairwise sufficiency of T and Lemma 12 imply that $d\mu_0/d(\mu_0 + \lambda)$ (ϵ) $T^{-1}(\boldsymbol{T})$. The relation

$$\frac{d\mu_0}{d\lambda} = \frac{d\mu_0}{d(\mu_0 + \lambda)} \cdot \frac{d(\mu_0 + \lambda)}{d\lambda} = \frac{d\mu_0}{d(\mu_0 + \lambda)} \cdot \left(\frac{d\lambda}{d(\mu_0 + \lambda)}\right)^{-1}$$
$$= \frac{d\mu_0}{d(\mu_0 + \lambda)} \left(1 - \frac{d\mu_0}{d(\mu_0 + \lambda)}\right)^{-1}$$

implies that $d\mu_0/d\lambda$ (ϵ) $T^{-1}(\boldsymbol{T})$; an application of Theorem 1 concludes the proof.

A comparison of Theorems 1 and 2 and Corollary 3 yields immediately the following consequence of Theorem 3.

COROLLARY 4. *A necessary and sufficient condition that the statistic* T *be sufficient for a dominated set* $\mathfrak{M}$ *of measures on* $\mathbf{S}$ *is that, for any two measures* μ *and* ν *in* $\mathfrak{M}$, $d\mu/d(\mu + \nu)$ (ϵ) $T^{-1}(\boldsymbol{T})$, *or, equivalently,* $d\mu \mid d\nu$ (ϵ) $T^{-1}(\boldsymbol{T})$.

10. The value of sufficient statistics in statistical methodology. We gather from conversations with some able and prominent mathematical statisticians that there is doubt and disagreement about just what a sufficient statistic is sufficient to do, and in particular about in what sense if any it contains "all the information in a sample." We therefore conclude this paper with a brief explanation of a point of view which, while not original with us, has not received due publicity.

Suppose a statistician $\mathcal{S}$ is to be shown an observation x drawn at random from some sample space $(X, \mathcal{S})$ on which an unknown measure, μ, of a set $\mathfrak{M}$ of possible measures obtains, while for the same observation x another statistician $\mathcal{T}$ is only to be shown the value $T(x)$ of some statistic T sufficient for $\mathfrak{M}$. It is clear that $\mathcal{S}$ is as well off as $\mathcal{T}$; we shall argue that $\mathcal{T}$ is also as well off as $\mathcal{S}$.

Suppose $\mathcal{S}$ has decided how to use his datum, that, in other words, he has decided just what he will do (or, in particular, say) in the event of each possible x. His program can then be described schematically by saying that he has selected some function f (of the points x) which, without serious loss of generality, may be supposed to take real values. Now $\mathcal{S}$'s only real concern is for the probability distribution of f given μ, i.e. for the function φ of a real variable c, defined by

$$\varphi(c) = \mu(\{x: f(x) < c\}) = \mu(E(c)).$$

But $\mathcal{T}$ can if he wishes achieve exactly the same results as $\mathcal{S}$, in the following way. Let him, on learning the value of $T(x)$, select a real number f, with the aid of a "random machine" which produces numerical values according to the known distribution function ψ, defined by

$$\psi(c) = p(E(c) \mid T(x)).$$

Then, for any μ in $\mathfrak{M}$, the probability that $\mathcal{T}$ will select a value less than c is

$$\int p(E(c) \mid y) d\mu T^{-1}(y) = \mu(E(c)) = \varphi(c).$$

Thus $\mathcal{T}$ is at no disadvantage, save for the mechanical one of having to manipulate a random machine, and he may fairly be said to have as much information as $\mathcal{S}$.

As a matter of fact we know of no practical situation in which $\mathcal{T}$ would actually go to the trouble of using a random machine. There are some situations in which he should in principle do so, but in which practical statisticians have not, so far as we know, thought it worth while. If, for example, an outcome consists of a sequence of n heads and tails resulting from n spins of a coin the heads ratio of which is known to be either one half or one quarter, then a sufficient statistic is the number of heads which occur in the sequence. In basing a decision on the outcome of this program both $\mathcal{S}$ and, to a still greater extent, $\mathcal{T}$ have (according to Wald's theory of minimum risk) something to gain by recourse to a random machine. There are, on the other hand, many technical desiderata which sufficient statistics meet exactly without recourse to random machines. Thus, as Blackwell has shown,[18] if $\mathcal{S}$ has an unbiased estimate, R, of some parameter, $\mathcal{T}$ can find a function R^*, defined by $R^*(y) = e(R \mid y)$, which is an unbiased estimate of that parameter, with variance not greater than that of R. More generally, if R is *any* estimate with finite mean square deviation from a parameter, then it is easy to show with Blackwell's methods that R^*

[18] D. Blackwell, "Conditional expectation and unbiased sequential estimation," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 105–110.

has no larger a mean square deviation than R. Finally it is a well known fact that, under suitable hypotheses, if there exists a maximum likelihood estimate R of some parameter, then R depends only on y.

We think that confusion has from time to time been thrown on the subject by (a) the unfortunate use of the term "sufficient estimate," (b) the undue emphasis on the factorability of sufficient statistics, and (c) the assumption that a sufficient statistic contains all the information in only the technical sense of "information" as measured by variance.