The Annals of Mathematical Statistics
1972, Vol. 43, No. 2, 490-497

STATISTICAL STRUCTURE OF THE PROBLEM OF SAMPLING FROM FINITE POPULATIONS

By Haruki Morimoto
Osaka City University

1. Introduction. Let X be a space, $\mathbf{A}$ a σ-field of subsets of X, and $P = \{p\}$ a family of probability measures on $\mathbf{A}$. A triplet $(X, \mathbf{A}, P)$ is often called a statistical structure. Basu and Ghosh [1] proposed as a measure-theoretic expression of the statistical problems of sampling from finite populations a special type of statistical structure satisfying the following assumptions.

Assumptions. (1) X is a space containing more than countably many points.
(2) $\mathbf{A}$ is the σ-field consisting of all subsets of X.
(3) Each $p \in P$ is a discrete probability measure.
(4) $p(A) = 0$ for all $p \in P$ implies $A = \emptyset$.

Section 2 of the present paper gives a few results about this structure. It is proved that a σ-field is inducible if and only if it is also a complete field, that is, it is closed under the formation of arbitrarily many (possibly more than countable) number of sets in it. It is also shown that the pairwise sufficiency of an inducible σ-field (a statistic) implies its sufficiency. Furthermore, the essential completeness of the class of tests which are measurable with respect to an inducible σ-field implies its sufficiency. It is observed that these results do not generally hold for non-inducible σ-fields. A partial analogue of Neyman's factorization theorem which characterizes pairwise-sufficient σ-fields is given.

Section 3 contains an attempt, based on the results in Section 2, to remove some inconveniences that are still remaining in the same structure.

2. Sufficiency and pairwise-sufficiency. We use here essentially the same definitions and similar notations as in [1]. Thus by a *statistic* we mean a *partition* $\mathbf{T} = \{T\}$, a class of mutually disjoint non-empty subsets of X which collectively cover X. Two partitions $\mathbf{T}$ and $\mathbf{U}$ are written $\mathbf{T} > \mathbf{U}$ when every set in $\mathbf{U}$ is a union of some sets in $\mathbf{T}$. A letter $\mathbf{B}$ will usually stand for a sub-σ-field of $\mathbf{A}$. Any statistic $\mathbf{T}$ induces a sub-σ-field $\mathbf{B}(\mathbf{T})$ of $\mathbf{A}$, the class of all those sets in $\mathbf{A}$ which can be expressed as a union of sets in $\mathbf{T}$. Two points y and z in X belonging to a common T in $\mathbf{T}$ are written $y \sim z(\mathbf{T})$. This is an equivalence relation on X, and conversely any equivalence relation gives rise to a statistic defined as the totality of its equivalence classes. On the other hand, $y \sim z(\mathbf{B})$ means that each set B in $\mathbf{B}$ either contains both or neither of y and z. This,

Received January 29, 1970; revised August 5, 1971.

too, is an equivalence relation and defines a partition $\mathbf{T}(\mathbf{B})$, which we call the *partition induced by* $\mathbf{B}$.

A sub-σ-field $\mathbf{B}$ is said to be *sufficient* (for $(X, \mathbf{A}, P)$), if for every $A \in \mathbf{A}$, there exists a $\mathbf{B}$-measurable function $f(x, A)$ such that, for all $B \in \mathbf{B}$ and $p \in P$,

$$p(A \cap B) = \int_B f(x, A)\, dp(x) .$$

$\mathbf{B}$ is said to be *pairwise sufficient* if it is sufficient for $(X, A, \{p, q\})$ for every p and q in P. A statistic $\mathbf{T}$ is *sufficient* if the induced sub-σ-field $\mathbf{B}(\mathbf{T})$ is sufficient. A sufficient σ-field is called *minimal sufficient* if it is contained in all other sufficient σ-fields. A sufficient statistic $\mathbf{T}$ is called *minimal sufficient* if for any sufficient statistic $\mathbf{U}$, $\mathbf{T} < \mathbf{U}$ holds.

$E[* \mid p]$ stands for the expectation of a function $*$ with respect to a measure p. For each p in P we define $A(p) = \{x \mid p(x) > 0\}$.

For later reference we summarize here the main results of [1] in the following

LEMMA 1. *Under the assumptions* (1)—(4),

(5) *A statistic* $\mathbf{T} = \{T\}$ *is sufficient if and only if each* $p \in P$ *is factored as*

$$p(x) = g(x, p)h(x) , \qquad x \in X ,$$

where $g(x, p)$ *is constant on each* T *and* $h(x)$ *is independent of* p.

(6) *The minimal sufficient statistic exists and induces the minimal sufficient* σ*-field.*

(7) *The minimal sufficient statistic* $\mathbf{M} = \{M\}$ *is constructed as follows;* $y \sim z(\mathbf{M})$ *if and only if* $p(y) > 0 \Leftrightarrow p(z) > 0$ *for all* p *in* P *and* $p(y)/p(z)$ *is independent of* p *for all* $p \in P$ *for which* $p(y)p(z) > 0$.

(8) *Every sufficient* σ*-field is induced by a sufficient statistic.*

(9) *For two statistics* $\mathbf{T}$ *and* $\mathbf{U}$, *if* $\mathbf{T} > \mathbf{U}$ *and if* $\mathbf{U}$ *is sufficient, then* $\mathbf{T}$ *is also sufficient.*

Now we give our first theorem which characterizes the inducible σ-fields. Note that (1)—(4) are assumed throughout this section.

THEOREM 1. *A* σ*-field* $\mathbf{B}$ *is inducible if and only if it is closed under the formation of unions of arbitrarily many (possibly more than countable) number of sets in it.*

PROOF. The "only if" part is clear from the definition of $\mathbf{B}(\mathbf{T})$.

To prove the "if" part we will show that $\mathbf{B}$ is induced by $\mathbf{T}(\mathbf{B})$, that is, $\mathbf{B}(\mathbf{T}(\mathbf{B})) = \mathbf{B}$. First, observe that every set T in $\mathbf{T}(\mathbf{B})$ belongs to $\mathbf{B}$, because T is the intersection of all those sets in $\mathbf{B}$ which contain T. From the assumption of arbitrary additivity of $\mathbf{B}$, it is also closed under the formation of intersection of arbitrarily many sets. Hence T belongs to $\mathbf{B}$.

Now let $A \in \mathbf{B}(\mathbf{T}(\mathbf{B}))$. It follows that A is a union of sets in $\mathbf{T}(\mathbf{B})$. But these

sets belong to **B**, as was shown in the last paragraph. Therefore $A \in \mathbf{B}$.

On the other hand, $A \in \mathbf{B}$ implies that A is a union of sets in $\mathbf{T}(\mathbf{B})$. But as $\mathbf{B}(\mathbf{T}(\mathbf{B}))$ contains all such sets, we see that $B \in \mathbf{B}(\mathbf{T}(\mathbf{B}))$. This completes our proof.

Hereafter we discuss some properties of sufficiency in this set-up. **M** denotes, as before, the minimal sufficient partition, whose existence was guaranteed by (6). Let **D** denote the σ-field consisting of all countable unions of sets in **M** and their complements.

THEOREM 2. *Let* **T** *be a sufficient statistic and* **B** *be the σ-field consisting of all countable unions of sets in* **T** *and their complements. Then* **B** *is pairwise sufficient. In particular,* **D** *is pairwise sufficient.*

PROOF. Each M in **M** is contained in some $A(p)$ (by (7)) and hence is a countable set. This and $\mathbf{T} > \mathbf{M}$ imply that M is a countable union of sets in **T**. Hence each D in **D** or its complement is a countable union of sets in **T**. Thus we have $\mathbf{B} \supset \mathbf{D}$, so that it is enough to prove the theorem for **D**.

Take any two measures p and q in P. Let $A = \{x \mid p(x)q(x) > 0\}$ and $B = \{x \mid p(x) > 0 \text{ and } q(x) = 0\}$. Define a discrete measure $m(x) = p(x) + q(x)$. Then we have

$$\begin{aligned} \frac{dp}{dm} &= \frac{1}{1 + (q(x)/p(x))}, && x \in A, \\ &= 1, && x \in B, \\ &= 0, && \text{otherwise}. \end{aligned}$$

A and B are unions of sets in **M** and, since $A \subset A(p)$ and $B \subset A(p)$, they are countable sets and hence in **D**. Therefore, $X - A - B$ is also in **D**. $q(x)/p(x)$ is constant on each $M \in \mathbf{M}$ such that $M \subset A$, by (7). As A is a countable union of sets in **M**, the set $\{x \mid dp/dm \leqq \alpha\}$ belongs to **D**, for all α such that $0 < \alpha < 1$. Thus dp/dm is **D**-measurable. Similarly dq/dm is **D**-measurable. Therefore, by the factorization theorem of Neyman **D** is sufficient for $\{p, q\}$.

REMARK 1. **D** is not sufficient for $(X, \mathbf{A}, P)$, since it does not contain any uncountable union of sets in **M** which are contained in $\mathbf{B}(\mathbf{M})$, the minimal sufficient σ-field for (X, A, P). This shows that the sufficiency and pairwise-sufficiency do not coincide with each other.

THEOREM 3. *Let* **B** *be pairwise sufficient for P. Then* $\mathbf{T}(\mathbf{B}) > \mathbf{T}(\mathbf{D}) = \mathbf{M}$.

PROOF. Take two points y and z in X such that $y \not\sim z(\mathbf{M})$. There exist p and q in P such that $q(y)/p(y) > q(z)/p(z)$, where $q(y)p(z) > 0$ although $p(y)$ may be 0. Put $\alpha = \sum \{p(x) \mid q(x)/p(x) > q(z)/p(z)\}$. Since **B** is sufficient for $\{p, q\}$ there exists a **B**-measurable most powerful test of p against q for the level α. From the fundamental lemma of Neyman and Pearson, the value of

this test function is 1 at y and 0 at z. Hence there must be some set in $\mathbf{B}$ which contains y but does not contain z. It means that $y \not\sim z(\mathbf{T}(\mathbf{B}))$.

COROLLARY 1. *A statistic is sufficient if and only if it is pairwise sufficient.*

PROOF. "Only if" part is true for any statistical structure. "If" part is shown as follows. From the pairwise sufficiency of $\mathbf{T}$, it follows that $\mathbf{T}(\mathbf{B}(\mathbf{T})) > \mathbf{T}(\mathbf{D}) = \mathbf{M}$. But $\mathbf{T} = \mathbf{T}(\mathbf{B}(\mathbf{T}))$, because $y \sim z(\mathbf{T}) \Leftrightarrow y \sim z(\mathbf{B}(\mathbf{T}))$ for any $\mathbf{T}$ and $y \sim z(\mathbf{B}) \Leftrightarrow y \sim z(\mathbf{T}(\mathbf{B}))$ for any $\mathbf{B}$ hold. So that $\mathbf{T} > \mathbf{M}$. Hence, by (9), $\mathbf{T}$ is sufficient.

THEOREM 4. Let a σ-field $\mathbf{B}$ contain $A(p)$ for all $p \in P$, and let $\mathbf{T}(\mathbf{B}) > \mathbf{T}(\mathbf{D}) = \mathbf{M}$. Then $\mathbf{B} \supset \mathbf{D}$ and $\mathbf{B}$ is pairwise sufficient.

PROOF. Take any T in $\mathbf{T}(\mathbf{B})$. $A(p) \in \mathbf{B}$ for all $p \in P$ implies that T is contained in some $A(p)$. $A(p)$, itself being countable, is a countable disjoint union of sets in $\mathbf{T}(\mathbf{B})$. Write $A(p) = T + T_1 + T_2 + \cdots + T_i + \cdots$, where $T_i \in \mathbf{T}(\mathbf{B})$, $i = 1, 2, \cdots$. For any i there exists a set B_i in $\mathbf{B}$ such that $B_i \supset T$ and $B_i \cap T_i = \emptyset$. It is easily seen that $T = (\bigcap_{i=1}^{\infty} B_i) \cap (A(p))$, so that $T \in \mathbf{B}$. Hence $\mathbf{T}(\mathbf{B}) \subset \mathbf{B}$.

Now take any set M in $\mathbf{M}$. From (7) we note that $M \subset A(p)$ for some $p \in P$ and is a countable set. This and the fact that $\mathbf{T}(\mathbf{B}) > \mathbf{M}$ imply that M is a countable union of sets in $\mathbf{T}(\mathbf{B})$. But $\mathbf{T}(\mathbf{B}) \subset \mathbf{B}$. Hence $M \in \mathbf{B}$. Thus $\mathbf{B} \supset \mathbf{M}$. But $\mathbf{D}$ is the smallest σ-field containing $\mathbf{M}$ so that $\mathbf{B} \supset \mathbf{D}$. Hence $\mathbf{B}$ is pairwise sufficient.

THEOREM 5. *Let $\mathbf{B}$ be a σ-field such that for any test function $t(x)$ there is a $\mathbf{B}$-measurable test function $s(x)$ with $E[t(x) \mid p] = E[s(x) \mid p]$ for all p in P. Then $\mathbf{T}(\mathbf{B}) > \mathbf{M}$ and $A(p) \in \mathbf{B}$ for all p in P. Consequently, $\mathbf{B}$ is pairwise sufficient.*

PROOF. First, for any $p \in P$, $A(p) = \{x \mid p(x) > 0\} \in \mathbf{B}$. To show this, take the indicator of $A(p)$ as $t(x)$. It then follows that $E[s(x) \mid p] = E[t(x) \mid p] = 1$. Hence $s(x) = 1$ in $A(p)$ and $s(x) \geqq t(x)$ for all $x \in X$. Therefore $s(x)$ has to coincide with $t(x)$ itself.

Next we will show that $\mathbf{T}(\mathbf{B}) > \mathbf{M}$. The proof is similar to that of Theorem 3. That is, letting $y \not\sim z(\mathbf{M})$ we construct a most powerful text $t(x)$ which, by our assumption, can be taken to be $\mathbf{B}$-measurable. But from the fundamental lemma of Neyman and Pearson $t(y) = 1$ and $t(z) = 0$. Hence $y \not\sim z(\mathbf{T}(\mathbf{B}))$.

REMARK 2. Pairwise sufficiency of $\mathbf{B}$ can also be proved directly from the assumption without invoking Theorem 4.

THEOREM 6. *Let $\mathbf{T} = \{T\}$ be a statistic such that for any test function $t(x)$ there exists a test function $s(x)$ which is constant on each $T \in \mathbf{T}$ and such that $E[t(x) \mid p] = E[s(x) \mid p]$ for all $p \in P$.*

Then T is sufficient for P.

PROOF. **B(T)** satisfies the assumption in Theorem 5, and hence is pairwise sufficient. Therefore **T** is pairwise sufficient, and is sufficient.

Finally we give a partial analogue of the factorization theorem (5) for general σ-fields.

THEOREM 7. *Let a σ-field* **B** *contain* **T(B)** *and let* **T(B)** *be sufficient. Then each p in P is factored as*

$$p(x) = g(x, p)h(x) \qquad x \in X,$$

such that $g(x, p)$ is **B***-measurable and $h(x)$ is independent of p.*

Conversely, if each p in P is factored as above, **B** *contains* **T(B)** *and* **T(B)** *is sufficient.*

PROOF. Sufficiency of **T(B)** implies that $p(x)$ is factored as in (5) where $g(x, p)$ is constant on each set in **T(B)** and $h(x)$ is independent of p. Since $g(x, p)$ vanishes outside the countable set $A(p)$, it is measurable with respect to any σ-field containing **T(B)**. Thus the first part of the theorem is proved.

The second part is shown as follows. From the **B**-measurability of $g(x, p)$, it is constant on each set in **T(B)**. Hence, by (5), **T(B)** is sufficient. Moreover, the set $A(p) = \{x \mid p(x) > 0\} = \{x \mid g(x, p) > 0\}$ belongs to **B** for all p in P. The first half of the proof of Theorem 4 now implies that $\mathbf{B} \supset \mathbf{T(B)}$.

REMARK 3. An example of an insufficient σ-field which satisfies the conditions of the foregoing theorem is given by **D**.

3. Discussion. We are concerned with a type of undominated statistical structure satisfying Assumptions (1)—(4), which hereafter will be referred to as "Basu-Ghosh's structure." Although it shares a number of convenient properties endowed by dominated structures as were listed in Lemma 1, there still remain some inconveniences involved by non-inducible σ-fields. In fact, Basu and Ghosh have already pointed out that no analogues of (9) hold true for non-inducible σ-fields. Some of the results in our Section 2 also emphasize this fact; for instance, compare (5) with Theorem 7, Corollary 1 with Remark 1, and Theorem 6 with Theorem 5.

Theorem 1 shows that inducible σ-fields are complete fields, i.e., closed under formation of unions of arbitrarily many (possibly uncountable number of) sets in it. Noting also that $\mathbf{A} = 2^X$ itself is a complete field, we propose now to re-write Basu-Ghosh's structure in the following form:

Statistical structure. We have a triplet $(X, \mathbf{A}, P)$ with the following properties.

(10) X is a space containing more than countable points.

(11) **A** is the complete field of all subsets of X.

(12) Each p in P is (not only a countably additive, but) a completely

additive probability measure. Namely, p is a nonnegative set-function on $\mathbf{A}$ satisfying

(12a) $$p(X) = 1 ,$$

and

(12b) $$p(\bigcup_{\gamma \in \Gamma} A_\gamma) = \sum_{\gamma \in \Gamma} p(A_\gamma) ,$$

where $\{A_\gamma \mid \gamma \in \Gamma\}$ is any (possibly uncountable) disjoint subfamily of $\mathbf{A}$.

(13) $p(A) = 0$ for all p in P implies $A = \emptyset$.

The right-hand side of (12b) means, as usual, $\sup \{\sum_{\delta \in \Delta} p(A_\delta) \mid \Delta$ is a finite subset of $\Gamma\}$. It is easily seen that a completely-additive probability measure on $\mathbf{A} = 2^X$ is a discrete probability measure on $\mathbf{A} = 2^X$.

Under this structure it would be natural that we treat only sub-complete-fields, and we need not and should not consider incomplete subfields. Thus we have a formulation of the sample survey problem which shares practically all the convenient properties of the dominated discrete case.

It must also be kept in mind that one should combine complete fields as such; two or more complete fields *generate* the minimal complete-field, rather than the σ-field, containing them. This distinguishes our formulation from a mere re-interpretation of Basu-Ghosh's structure, as will be illustrated in the sequel.

Let us be given two finite populations from each of which a sample is drawn. Suppose that we take two structures of Basu-Ghosh's type, say $(X, 2^X, P)$ and $(Y, 2^Y, Q)$, each of which expresses separately the sampling from each of two populations. To set up a structure $(Z, \mathbf{C}, R)$ for the entire problem of simultaneous sampling from two populations, statisticians will usually form a direct product: $(Z, \mathbf{C}, R) = (X \times Y, 2^X \times 2^Y, P \times Q)$. Here, $2^X \times 2^Y$ is the minimal σ-field containing the totality of rectangles; $\{A \times B \mid A \subset X$ and $B \subset Y\}$; and $P \times Q = \{p \times q \mid p \in P$ and $q \in Q\}$. But on the other hand, as the sample space Z is $X \times Y$, Basu-Ghosh's argument would suggest putting $\mathbf{C} = 2^{X \times Y}$, the family of all subsets of $X \times Y$. However, the question as to whether

(14) $$2^{X \times Y} = 2^X \times 2^Y$$

holds true or not is not quite an obvious one, known as the "rectangle problem."

Various cases turn up, depending on the cardinalities of X and Y. If $\overline{\overline{X}}$ and $\overline{\overline{Y}}$ are both greater than $2^{\aleph_0}$, the answer to this question is negative. (Personal communication from M. Takahashi.) The case of the most practical importance is naturally the one where $\overline{\overline{X}}$ and/or $\overline{\overline{Y}} = 2^{\aleph_0}$, about which the following results seem to be worth mentioning. (Personal communication from K. Namba. Refer also to [3].)

Under Zermelo-Fraenkel's axioms of set theory with the axiom of choice

and the continuum hypothesis, (14) is proved affirmatively. However, if we adopt another hypothesis

$$2^{\aleph_0} = 2^{\aleph_1} = \aleph_2 , \tag{15}$$

instead of the continuum hypothesis, there exists a model in which at least one subset of $2^{\aleph_0} \times 2^{\aleph_0}$ is not in the σ-field generated by its rectangles. Hence, when $X = Y = R$, neither (14) nor

$$2^{X \times Y} \neq 2^X \times 2^Y \tag{16}$$

can be proved in Zermelo-Fraenkel's set theory with the axiom of choice.

Thus we cannot assume (14) too easily. Now we will see what happens under (16). In such cases the σ-field $2^{X \times Y}$ is not the direct product σ-field of 2^X and 2^Y. After having defined the direct product measure $p \times q$ on the direct product σ-field $2^X \times 2^Y$, we have to extend it again to $2^{X \times Y}$.

Under (16), $2^X \times 2^Y$, as a sub-σ-field of $2^{X \times Y}$, is not sufficient for $(X \times Y, 2^{X \times Y}, P \times Q)$. Because, $2^X \times 2^Y$ clearly induces the pointwise partition on $X \times Y$, that is to say, $\mathbf{T}(2^{X \times Y}) = \mathbf{T}(2^X \times 2^Y)$. Hence it is a non-inducible σ-field. But by (8), it has to be inducible so as to be sufficient. Therefore it is not sufficient. This fact serves as an example showing that the direct product of two σ-fields, one of which is sufficient for $(X, 2^X, P)$ and another for $(Y, 2^Y, Q)$, may not be sufficient for $(X \times Y, 2^{X \times Y}, P \times Q)$.

All of these complications are consequences of a σ-field $2^X \times 2^Y$ being understood as the direct product of 2^X and 2^Y. On the other hand, the smallest complete field generated by the totality of rectangles is clearly $2^{X \times Y}$. Therefore we are justified in taking $2^{X \times Y}$ as the direct product of 2^X and 2^Y when both of the latter are looked upon as complete-fields.

One can easily rewrite usual measure-theoretical results about statistical structures with discrete measures in terms of complete fields and completely-additive measures instead of σ-fields and countably-additive measures, respectively. In fact, the essential part of it has already been developed by Blackwell and Girschick [2]. They take an arbitrary sample space X and a family of discrete measures $P = \{p\}$. For any subset A of X, $p(A)$ is defined by $p(A) = \sum \{p(x) \mid x \in A \text{ and } p(x) > 0\}$. It clearly coincides with (10)—(13), though they do not (or rather need not!) treat σ-fields or complete-fields. They even proceed to prove Neyman's factorization theorem and the existence of a minimal sufficient statistic under this scheme.

4. Acknowledgments. The author is grateful to Professor H. Kudo, Messrs. M. Sibuya, K. Namba and M. Takahashi for many useful suggestions they gave to him. He is also thankful to referees for their suggestions.

REFERENCES

[1] BASU, D. and GHOSH, J. K. (1969). Sufficient statistics in sampling from a finite universe. *Proc. 36th session Internat. Statist. Inst.* 850–859.

[2] BLACKWELL, D. and GIRSCHICK, M. A. (1954). *Theory of Games and Statistical Decisions.* Wiley, New York.

[3] KUNEN, A. (1968). Item T3221 in "A survey of recent results in set theory," compiled and circulated by A.R.D. Mathias (mimeographed preprint).