GENERALIZED RANDOM VARIABLES

JOHN G. KEMENY

We will consider random variables on a denumerably infinite sample space. However, the range $\boldsymbol{R}$ of the random variables will not necessarily be a set of real numbers. In Part I the range will be a subset of a given metric space, and in Part II it will be an arbitrary set. Since each distribution on the sample space determines a distribution on $\boldsymbol{R}$ (for a given random variable), the sample space may be ignored entirely, and we may restrict our attention to distributions on $\boldsymbol{R}$. Thus, instead of discussing means and variances of random variables on the sample space, we will discuss means and variances of distributions on the set $\boldsymbol{R}$.

In classical probability theory $\boldsymbol{R}$ would be a set of real numbers, and the mean and variance of a distribution on $\boldsymbol{R}$ would also be real numbers. Of these restrictions only one will be kept, namely that the variance will always be a non-negative real number. As indicated above, $\boldsymbol{R}$ may be a more general space, and the means will also be selected from more general spaces. The defining property of a mean will be the property of minimizing the variance of the given distribution. It will be shown that these means still have many of the classical properties, though in general means are not unique, and in certain circumstances there may be no mean.

While the mean is classically taken to be a real number, it need not be an element of $\boldsymbol{R}$. For example, the mean of a set of integers may be a fraction. This approach is extended in Part I, where the means may be arbitrary points of a certain metric space $\boldsymbol{T}$, and $\boldsymbol{R}$ is any subset of $\boldsymbol{T}$. Even the form chosen for the variance is the same as in classical probability theory.

In Part II the concept of a random variable and of means is further generalized. Here $\boldsymbol{R}$ is an arbitrary set, and the topological space $\boldsymbol{T}$ from which means are chosen need not be metric and need bear no relation to $\boldsymbol{R}$. The variance is still a numerical function on $\boldsymbol{T}$, but of a much more general form than in Part I. In both frameworks an analogue of the strong law of large numbers is proved, to show that classical results can be generalized to these new kinds of random variables.

In Part III we consider certain generalizations. The positive result in this part is that the restriction to independent random variables in Parts I and II is unnecessary; the results hold for any metrically

Received February 11, 1959. The author wishes to acknowledge the aid of the National Science Foundation through a grant given to the Dartmouth Mathematics Project. He also wishes to thank H. Mirkil for a number of valuable suggestions.

transitive stationary process. There are also certain negative results, showing that some "obvious generalizations" fail.

PART I

We consider a metric space $\boldsymbol{T}$, from which our means will be selected. Let φ be the metric on $\boldsymbol{T}$. The range $\boldsymbol{R} = \{r_i\}$ of our random variables may be any denumerable subset of $\boldsymbol{T}$. We impose one restriction on the space $\boldsymbol{T}$:

(1) The closed spheres in $\boldsymbol{T}$ are compact.

As indicated above, instead of considering the random variables themselves, we will consider only distributions $P = \{p_i\}$ on $\boldsymbol{R}$. For each such distribution we define two numerical functions on $\boldsymbol{T}$:

$$W_P(t) = \sum_i \varphi(r_i, t) \cdot p_i \text{ and } V_P(t) = \sum_i \varphi^2(r_i, t) \cdot p_i .$$

The former may be thought of as a mean distance to t with respect to P, and the latter as a variance computed with respect to t. Both functions are non-negative real valued, with $+\infty$ as a possible value. We will, however, consider only distributions that satisfy the condition:

(2) There is a $t_0 \in \boldsymbol{T}$ such that $V_P(t_0)$ is finite.

DEFINITION. The *variance* of the distribution P is defined as

$v_P = \inf_{t \in \boldsymbol{T}} V_P(t)$. An element t of $\boldsymbol{T}$ is a *mean* of P if $V_P(t) = v_P$. We denote the set of means of P by M_P.

In view of this definition we note that (2) is equivalent to the assumption that P has finite variance. Should P have infinite variance, then all points of $\boldsymbol{T}$ would be means of P according to the definition. While the theorems to be proven below would all be true, they would become trivial. It is possible to give a more sophisticated definition of the mean for the case of an infinite variance (see [3]), but this leads into problems beyond the scope of the present paper.

LEMMA 1. *If $W_P(t_1)$ and $V_P(t_1)$ are finite, then*

$$| W_P(t_1) - W_P(t_2)| \leq \varphi(t_1, t_2) \text{ and}$$
$$| V_P(t_1) - V_P(t_2)| \leq \varphi(t_1, t_2) \cdot [2V_P(t_1) + 2 + \varphi(t_1, t_2)] .$$

Proof. $$| W_P(t_1) - W_P(t_2)| \leq \sum_i |\varphi(r_i, t_1) - \varphi(r_i, t_2)| \cdot p_i \leq \sum_i \varphi(t_1, t_2) \cdot p_i = \varphi(t_1, t_2)$$

using the triangle inequality on φ.

$$\begin{aligned} |V_P(t_1) - V_P(t_2)| &\le \sum_i |\varphi^2(r_i, t_1) - \varphi^2(r_i, t_2)| \cdot p_i \\ &\le \varphi(t_1, t_2) \cdot \sum_i [\varphi(r_i, t_1) + \varphi(r_i, t_2)] \cdot p_i \\ &= \varphi(t_1, t_2) \cdot [W_P(t_1) + W_P(t_2)] \\ &\le \varphi(t_1, t_2) \cdot [2W_P(t_1) + \varphi(t_1, t_2)] \end{aligned}$$

where in the second step we factored and applied the triangle inequality, while in the last step we used the result proved above. The second part of the lemma then follows if we observe that $W_P(t) \le V_P(t) + 1$ for any t.

LEMMA 2. *$W_P(t)$ and $V_P(t)$ are finite for all t.*

Proof. This is a consequence of the restriction (2). Choose t_0 as in (2). Then $W_P(t_0)$ is also finite. Lemma 1 yields that $|V_P(t_0) - V_P(t)|$ is finite, hence the lemma follows.

LEMMA 3. *$V_P(t)$ is a continuous function of t.*

Proof. Suppose that $t_1, t_2, \cdots$ is a sequence converging to t; then by Lemma 1, $|V_P(t) - V_P(t_k)| \le \varphi(t, t_k) \cdot [2V_P(t) + 2 + \varphi(t, t_k)]$. But $\varphi(t, t_k) \to \varphi(t, t) = 0$, hence $V_P(t_k) \to V_P(t)$.

There is one closed sphere that will occur frequently below. Let $S_1 = \text{Sph}\,(\sqrt{6V_P(r_1)/p_1}, r_1)$, that is the set of all points in $\boldsymbol{T}$ whose φ-distance from r_1 is at most the specified amount. Then S_1 is compact by (1).

THEOREM I. *M_P is a non-empty compact set.*

Proof. If $t \notin S_1$, then $V_P(t) \ge \varphi^2(r_1, t) \cdot p_1 > 6V_P(r_1)$. Hence $V_P(t)$ is bounded away from v_P. Thus the inf of V_P on all of T is the same as on S_1. But V_P is continuous, by Lemma 3, and hence $V_P(S_1)$ is compact. This means that V_P actually takes on its inf on S_1, hence P has at least one mean. Furthermore $M_P = V_P^{-1}(\{v_p\})$, hence it is a closed subset of S_1, and thus compact.

We will suppose that a sequence of point $x_1, x_2, \cdots$ is selected from $\boldsymbol{R}$. The points are selected independently, at random, according to a distribution Q satisfying (2). [We may consider $\boldsymbol{R}$ to be our new sample space, and the x_j to be identity functions on $\boldsymbol{R}$. Then they are independent, identically distributed generalized random variables.] For each n, we associate a distribution $H^n = \{h_i^h\}$ with the first n points in this sequence; namely h_i^n is the fraction of the first n points that are equal to r_i, or h_i^n is the frequency of occurrence of r_i among the first n random variables. It will be convenient to write V_n in place of

V_{H^n}, to write v_n for the variance of H^n, and M_n for the set of means of H^n. Clearly H^n has the property (2).

LEMMA 4. *For any* $t \in T$, $V_n(t) \to V_Q(t)$ *with probability* 1.

Proof. We may consider $\varphi^2(x_j, t)$, for a fixed t, to be a sequence of ordinary random variables on $\boldsymbol{R}$. They are independent and identically distributed. Their mean is $V_Q(t)$. Since this is finite by Lemma 2, the ordinary strong law of large numbers applies to them. But this states precisely that $V_n(t) \to V_Q(t)$ with probability 1. (See [2], p. 208).

LEMMA 5. *For any compact set* C *there is probability* 1 *that* $V_n(t) \to V_Q(t)$ *uniformly on* C.

Proof. Since C is compact and V_Q is continuous, $V_Q \leq A$ on C. For any integer k we can find a finite set of points C_k such that the spheres of radius $1/k$ about points in C_k cover C. Since the union of the sets C_k is denumerable, it follows from Lemma 4 that there is probability 1 that $V_n(t) \to V_Q(t)$ for all points in all the C_k. [The set of sequences on which convergence fails at one point has measure 0, hence the union of all these denumerably many sequences has measure 0, and hence the complement of the union has measure 1.] We restrict ourselves to such sequences of x_j. Let t be any point in C. Select a point $t_k \in C_k$ so that $\varphi(t, t_k) < 1/k$. Then

$$\begin{aligned}
| V_n(t) - V_Q(t) | &\leq | V_n(t) - V_n(t_k) | + | V_n(t_k) - V_Q(t_k) | \\
&\qquad + | V_Q(t_k) - V_Q(t) | \\
&\leq \frac{1}{k}\left[2V_n(t_k) + 2 + \frac{1}{k} \right] + | V_n(t_k) - V_Q(t_k) | \\
&\qquad + \frac{1}{k}\left[2V_Q(t_k) + 2 + \frac{1}{k} \right] \\
&\leq \frac{2}{k}[2A + 3] + 3 | V_n(t_k) - V_Q(t_k) |
\end{aligned}$$

where Lemma 1 was used in step 2, and the uniform bound A applied and terms combined in step 3.

Given $\varepsilon > 0$, we choose k large enough to make the first term less than $\varepsilon/2$. Since C_k has only a finite number of elements, for sufficiently large n, $| V_n(t_k) - V_Q(t_k) | < \varepsilon/2$ for all $t_k \in C_k$. Hence for sufficiency large n, $| V_n(t) - V_Q(t) | < \varepsilon$ for all $t \in C$.

LEMMA 6.[1] *With probability* 1, *for sufficiently large* n $M_n \subseteq S_1$.

Proof. By the ordinary strong law of large numbers, $h_1^n \to q_1$ with probability 1; and by Lemma 4, $V_n(r_1) \to V_Q(r_1)$ with probability 1. Hence we can select a sequence with probability 1 on which both events take place. On such a sequence, for sufficiently large n,

$$v_n \leq V_n(r_1) < 2V_Q(r_1) \quad \text{and} \quad h_1^n > q_1/2 \; .$$

Hence, if $t \notin S_1$, then for sufficiently large n,

$$V_n(t) \geq \varphi^2(r_1, t) \cdot h_1^n > (6V_Q(r_1)/q_1) \cdot (q_1/2) = 3V_Q(r_1) \; ,$$

which is bounded away from v_n. Hence if $t \notin S_1$, then $t \notin M_n$. And hence $M_n \subseteq S_1$.

We are now in a position to prove a version of the strong law of large numbers. This states that the sequence of sample means converges to the mean of the distribution with probability 1. In our more general framework we do not have unique means, though we do have assurance from Theorem I that the set of means is non-empty both for the samples and for the distribution. We thus want to prove that the sequence of sets M_n converges to the fixed set M_Q with probability 1. As the criterion for convergence we require that every open set containing M_Q should contain almost all M_n. If the means happen to be unique, this is equivalent to ordinary convergence.

THEOREM II. $M_n \to M_Q$ *with probability* 1.

Proof. By Lemma 5, there is probability 1 that $V_n(t) \to V_Q(t)$ uniformly on S_1. By Lemma 6, almost all M_n are subsets of S_1 with probability 1. Hence with probability 1 we may restrict ourselves to x_j-sequences on which both events occur. Let O be an open set containing M_Q. Then $S_1 \cap \tilde{O}$ is compact, and hence V_Q takes on a minimum value v on it. But no mean is in this set, hence $v > v_Q$. Let $m \in M_Q$ and $t \in S_1 \cap \tilde{O}$.

$$V_n(t) - V_n(m) = [V_n(t) - V_Q(t)] + [V_Q(t) - V_Q(m)] + [V_Q(m) - V_n(m)] \; .$$

From the uniform convergence of V_n we know that for sufficiently large n the first and third terms will both be less than $(v - v_Q)/3$ in absolute value. The middle term is at least $v - v_Q$. Hence for sufficiently large n the difference is positive, and hence $t \notin M_n$. Hence no element of $S_1 \cap \tilde{O}$ is in M_n, and we also know that $M_n \subseteq S_1$ for almost all n. Hence $M_n \subseteq O$ for almost all n.

[1] S_1 is here defined with respect to the Q-distribution, that is, V_Q and q_1 take the place of V_P and p_1 in the definition. This will be the sphere used from here on in Part I.

THEOREM III. *$v_n \to v_Q$ with probability* 1.

Proof. $|v_n - v_Q| \leq |V_n(m_n) - V_Q(m_n)| + |V_Q(m_n) - V_Q(m)|$ if $m_n \in M_n$, $m \in M$.

As in the previous theorem, we may combine Lemmas 5 and 6 to assure that the first term tends to 0 with probability 1. The sequence of m_n's will, with probability 1, have a limit point, by Lemma 6 and the compactness of S_1. And by Theorem II this limit point will, with probability 1, be in M_Q. It then follows from the continuity of V_Q that if we choose as m this limit point, the second term goes to 0 with probability 1.

One interesting set of applications of these theorems may be obtained by choosing for $\boldsymbol{T}$ a metric space with compact spheres, and choosing for φ a suitable function of the metric. If d is the metric, and f is a numerical function such that $f(0)=0, f'>0$ and $f'' \leq 0$, then $\varphi(t_1, t_2) = f(d(t_1, t_2))$ is also a metric on $\boldsymbol{T}$. In particular, we may choose $\varphi = d^k$, for $k \leq 1$. The choice of $k=1$ yields the generalization of the ordinary arithmetic mean, and $k = 1/2$ yields a generalization of the median.

If for $\boldsymbol{T}$ we choose Euclidean n-space, and let $\varphi = d$, then Theorem II yields the classical strong law of large numbers for the case of discrete random variables with a finite variance.

Condition (1) is a natural condition to impose when generalizing results from Euclidean n-space. But it is reasonable to ask whether the condition is really necessary. For example, could one replace it by the assumption that $\boldsymbol{T}$ is locally compact? The following example shows that local compactness does not suffice: Let $\boldsymbol{T} = \boldsymbol{R} \cup \boldsymbol{S}$, with $\boldsymbol{S} = \{s_i\}$ for $i = 1, 2, \cdots$. We introduce the metric φ as follows.

$$\varphi(r_i, r_j) = \varphi(s_i, s_j) = 2(1-\delta_{ij}) \quad \text{and} \quad \varphi(r_i, s_j) = \begin{cases} 1 & \text{if } j \geq i \\ 2 & \text{if } j < i \end{cases}.$$

Let $p_i = 1/2^i$. Then $v_P = 1$, and r_1 is the unique mean. Suppose that H^n is a close approximation of P, with $h_1^n \leq 1/2$. This has positive probability. If i_0 is the last i for which $h_1^n > 0$, then s_j is a mean of H^n for all $j \geq i_0$. Hence M_n does not converge to $M_P = \{r_1\}$. This metric topology, which happens to be discrete, violates condition (1), but $\boldsymbol{T}$ is locally compact.

PART II

We will now consider a more general framework in which $\boldsymbol{R}$ is an arbitrary set, and $\boldsymbol{T}$ any topological space. We will consider the space $\boldsymbol{P}$ of all possible measures $P = \{p_i\}$ on $\boldsymbol{R}$. But since $\boldsymbol{R}$ is an arbitrary denumerable infinite set, we may—without loss of generality—take P to be a measure on the integers. The basic tool in Part I was a numerical function $V_P(t)$ on $\boldsymbol{T}$, for each measure $\boldsymbol{P}$, satisfying certain conditions.

We will again assume that there is a function V_P corresponding to each P in $\boldsymbol{P}$.

We will introduce metrics on two basic spaces. On the space $\boldsymbol{P}$ of all measures on the integers we define $d(P, Q) = \sum_i |p_i - q_i|$. On the space $\boldsymbol{F}$ of all non-negative bounded real-valued functions on $\boldsymbol{T}$ we define $d(f, g) = \sup_{t \in \boldsymbol{T}} |f(t) - g(t)|$.

Our basic assumptions concern the mapping $P \rightarrow V_P$ from $\boldsymbol{P}$ into $\boldsymbol{F}$. We require that:

(1) Each image V_P is a function that takes on a minimum on every closed subset of $\boldsymbol{T}$.

(2) The mapping is continuous.

We may then introduce means and variances as in the definition in Part I. We may prove near analogues of the previous theorems.

THEOREM I′. *M_P is non-empty for each $P \in \boldsymbol{P}$.*

Proof. V_P takes on a minimum on every closed subset of $\boldsymbol{T}$, by (1), hence it takes on its minimum on $\boldsymbol{T}$.

We will again consider sequences x_j, selected independently at random according to a distribution Q. We define the sample distributions, means, and variances as in Part I.

LEMMA. *$H^n \rightarrow Q$ with probability* 1.

Proof. The lemma asserts that $d(H^n, Q) \rightarrow 0$ with probability 1. From the definition of the metric on P we see that this asserts that $\sum_i |h_i^n - q_i| \rightarrow 0$ with probability 1. This was proved by Parzen in a paper that has not yet appeared (see [4]).

THEOREM II′. *$M_n \rightarrow M_Q$ with probability* 1.

Proof. Let O be an open set containing M_Q. Then $\tilde{O}$ is closed, and hence V_Q takes on a minimum value on it, by (1), say v. Since no mean of Q is in $\tilde{O}$, $v > v_Q$.

Suppose that $H^n \rightarrow Q$ in $\boldsymbol{P}$, which occurs with probability 1 by the lemma. Then by (2), $V_n \rightarrow V_Q$ in $\boldsymbol{F}$. But this means that $V_n(t)$ converges uniformly to $V_Q(t)$. Let $t \in \tilde{O}$, $m \in M_Q$, then

$$V_n(t) - V_n(m) = [V_n(t) - V_Q(t)] + [V_Q(t) - V_Q(m)] + [V_Q(m) - V_n(m)] .$$

By the uniform convergence of V_n we can make the first and third terms less in absolute value than $(v - v_Q)/3$, for all $t \in \tilde{O}$, for sufficiently large n. The middle term is at least $v - v_Q$, hence for all sufficiently large n the difference is positive, and hence for these n, $M_n \subseteq O$.

THEOREM III′. *$v_n \to v_Q$ with probability* 1.

Proof. $v_n \leq V_n(m) \leq v_Q + |V_n(m) - V_Q(m)|$, for $m \in M_Q$.

And $v_Q \leq V_Q(m_n) \leq v_n + |V_Q(m_n) - V_n(m_n)|$, for $m_n \in M_n$.
Hence, $|v_n - v_Q| \leq \sup_{t \in T} |V_n(t) - V_Q(t)| = d(V_n, V_Q)$.

But this tends to 0 with probability 1, by the lemma and (2).

Let us consider some applications of these theorems. First we will suppose that $V_P(t) = \Sigma \varphi^2(r_i, t) \cdot p_i$, where φ is a numerical function on $\boldsymbol{R} \times \boldsymbol{T}$. This is the nearest analogue we have to Part I. But even in this case the assumptions made in Part II are not comparable to those in Part I. The easiest way to assure that (2) is satisfied is to require that $|\varphi| \leq B$ on $\boldsymbol{R} \times \boldsymbol{T}$. Then V_p is always bounded, and

$$|V_P(t) - V_Q(t)| \leq \sum_i \varphi^2(r_i, t) \cdot |p_i - q_i| \leq B^2 \sum_i |p_i - q_i| .$$

Hence $d(V_P, V_Q) \leq B^2 \cdot d(P, Q)$. Hence the mapping $P \to V_P$ is continuous.

There are various ways of fulfilling (1). One very interesting case is where $\boldsymbol{T}$ is compact and $\varphi(r_i, t)$ is lower semi-continuous on $\boldsymbol{T}$ for each $r_i \in \boldsymbol{R}$. Then every closed subset is compact, and hence a lower semi-continuous function will take on its minimum on it. And V_P is the uniform limit of a sequence of monotone increasing lower semi-continuous functions, hence it itself has this property.

Thus if $\boldsymbol{T}$ is compact, we may choose as φ any function bounded on $\boldsymbol{R} \times \boldsymbol{T}$, such that each $\varphi(r_i, t)$ is lower semi-continuous on $\boldsymbol{T}$. Obvious examples of this may be found by choosing $\boldsymbol{R} \subseteq \boldsymbol{T}$, where $\boldsymbol{T}$ is a compact metric space and φ a continuous function of the distance. Thus we see that if we are willing to assume that $\boldsymbol{T}$ is compact, we are allowed to choose φ in much greater generality than in Part I.

If, in particular, $\boldsymbol{T}$ is a finite metric space, then Theorem II′ has an interesting corollary. Since the topology is discrete in this case, $M_n \to M_Q$ implies that $M_n \subseteq M_Q$ for sufficiently large n. Hence there is probability 1 that for sufficiently long sample sequences all sample means are means of the distribution. If the distribution has a unique mean, then there is probability 1 that all sufficiently large samples have this mean as their unique mean.[2]

Let us now consider an example of a compact space with a bounded φ, where φ is only lower semi-continuous. Let $\boldsymbol{T}$ be the set of all vectors $\{a_i\}$, $i = 1, 2, \cdots$, where $a_i \geq 0$ and $\sum_i a_i \leq 1$. We define the distance $\varphi(A, B)$ between two vectors as $\sum_i |a_i - b_i|$. However, $\boldsymbol{T}$ is

[2] An interesting application of this result is worked out by the author and J. L. Snell in a forthcoming book: It can be shown that there is a "natural" metric for the space of all rankings of k individuals. Thus our result allows certain statistical procedures for rankings.

not compact with respect to the metric topology. So we choose for $\boldsymbol{T}$ a weaker topology, namely the topology of componentwise convergence of vectors. $\boldsymbol{T}$ is compact with respect to this weaker topology, and φ is lower semi-continuous on the resulting topological space. Clearly, $\varphi \leq 2$, hence all our conditions are satisfied, and hence the theorems are applicable.

Let us next consider an example where φ is bounded, but $\boldsymbol{T}$ is not compact. Let $\boldsymbol{R} \subseteq \boldsymbol{T}$, and $\boldsymbol{T}$ an arbitrary topological space. Define $\varphi(r, t) = 1 - \delta_{rt}$. Then $V_P(t) = 1$ if $t \notin \boldsymbol{R}$, and $1 - p_i$ if $t = r_i$. Hence on a set not intersecting $\boldsymbol{R}$ the minimum of 1 is taken on at all points, while otherwise the minimum is taken on where p_i is largest. Hence V_P satisfies condition (1), and we see that M_P is always a non-empty, finite set. It is the set of *modes* of the distribution P.

Finally, let us consider one example where V_P is not of the general form of Part I. Let $\boldsymbol{T}$ be a compact metric space, and $\boldsymbol{R} \subseteq \boldsymbol{T}$. If d is the metric, we define $V_P(t)$ as the inf of $\sum_{i \in J} d(r_i, t) \cdot p_i$ over all sets J such that $\sum_{i \in J} p_i > .9$. Here V_P is lower semi-continuous on the compact set, hence (1) holds. Since d is bounded on $\boldsymbol{T}$, say by B, a change of ε in P will produce a change of at most $B\varepsilon$ in V_P; hence (2) holds.

This result has the following "practical" application. Suppose that a state legislature decides to establish a state university. They may insist that the University service at least 90 percent of the state's population, and that it be in the "most convenient location" for the population. This may be interpreted by introducing as a metric distance between r_i and t the distance a person at r_i has to travel to reach a university located at t. Then we find the mean of the V_P described above, with p_i taken proportional to the population at location r_i. Theorem II′ then states that if the college population is a cross-section of the entire population, and if the university is large enough, then there is an excellent chance that the location "most convenient for the entire population" will be "most convenient for the freshman class" in any given year. While the practicality of this procedure is debatable, it is more reasonable than the location of a state university in the geometric center of the state. It also shows that the theorems of Part II lend themselves to many unorthodox applications.

It is worth remarking that if V_P is lower semi-continuous, then M_P will be closed. So in all the examples discussed above where V_P was a lower semi-continuous function on a compact space, we obtain the full equivalent of Theorem I, since M_P is compact. And in the example of the modes M_P was finite.

But we can't always expect this to happen in the very general framework of Part II. As a matter of fact, M_P may be any subset of

$\boldsymbol{T}$. Let S be a subset of $\boldsymbol{T}$, and define $f(t)$ to be 0 on S and 1 on $\tilde{S}$. If we assign this same function to all distributions, that is $V_P = f$ for all P, then both conditions (1) and (2) are fulfilled. But $M_P = S$.

It may be worth pointing out that the mapping $P \rightarrow V_P$ need not be defined on all of $\boldsymbol{P}$. It suffices if it is defined and continuous on a subspace, as long as this subspace includes all measures having only a finite number of positive p_i. The theorems then apply to measures in the subspace. This extra freedom is convenient in a situation where the desired definition of V_P leads to unbounded functions for certain distributions.

PART III

In conclusion we will show that certain other classical ideas fail to generalize. If X_1, X_2 are two random variables, we can introduce a *mean random variable* X which corresponds to $1/2(X_1 + X_2)$ in the classical case. We define the value of X to be the mean of the values of X_1, X_2, if there is a unique mean. If there is more than one mean, we assume that X is equally likely to take on each of these values. We would at least expect that if X_1 and X_2 have the same unique mean, then X also has this mean. However, Figure 1 shows a distribution on a metric space with eight points (each line represents a unit distance), which provides a counter-example. If X_1 and X_2 each have the distribution of Figure 1, then X has the distribution of Figure 2. While X_1 and X_2 have the unique mean A, X has the unique mean B.

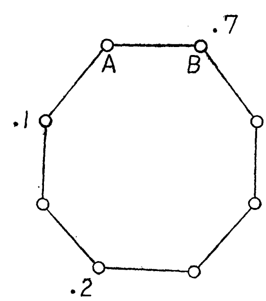

Fig. 1.

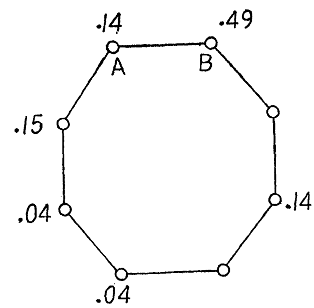

Fig. 2.

Next we will consider classical proofs using Chebyshev's inequality. We may state a version of this inequality, in the terminology of Part I as

$$\Pr\left[\varphi(x, M_Q) > k\right] \leq v_Q/k^2 \ .$$

This inequality may be proved by an exact analogue of the classical argument. However, the usual method for obtaining the weak law of large numbers from it fails. We would need to show that if we define a mean random variable X, as above, its variance tends to 0. However, this is rarely the case. If, for example, R consists of two points, and we have probability 1/2 for each point, then the variance of X tends to 1/2.

On the other hand, it is easy to extend our results to stochastic processes more general than those considered so far. In Parts I and II only identically distributed independent generalized random variables were considered. However, the only property of the process used in Part I was that the strong law of large numbers held. In Part II only the Parzen result was used. Both of these hold for metrically transitive stationary processes (see [1], Ch. X Sects. 1–2, and [4]). Hence all our results hold for these stochastic processes.

BIBLIOGRAPHY

1. J. L. Doob, *Stochastic processes*, John Wiley and Sons, 1953.

2. W. Feller, *Probability theory and it applications*, John Wiley and Sons 1950.

3. M. Fréchet, *Les éléments aléatoires de nature quelconque dans un espace distancié*, Institut Henri Poincaré, Vol. X, No. IV, 1948

4. E. Parzen, *On the convergence of distribution functions at points of discontinuity* forthcoming.

DARTMOUTH COLLEGE