THEORY OF PROBABILITY AND ITS APPLICATIONS

Volume XVI | Number 2

1971

ON THE UNIFORM CONVERGENCE OF RELATIVE FREQUENCIES OF EVENTS TO THEIR PROBABILITIES

V. N. VAPNIK AND A. YA. CHERVONENKIS

(*Translated by B. Seckler*)

Introduction

According to the classical Bernoulli theorem, the relative frequency of an event A in a sequence of independent trials converges (in probability) to the probability of that event. In many applications, however, the need arises to judge simultaneously the probabilities of events of an entire class S from one and the same sample. Moreover, it is required that the relative frequency of the events converge to the probability uniformly over the entire class of events S. More precisely, it is required that the probability that the maximum difference (over the class) between the relative frequency and the probability exceed a given arbitrarily small positive constant should tend to zero as the number of trials is increased indefinitely. It turns out that even in the simplest of examples this sort of uniform convergence need not hold. Therefore, one would like to have criteria on the basis of which one could judge whether there is such convergence or not.

This paper first indicates sufficient conditions for such uniform convergence which do not depend on the distribution properties and furnishes an estimate for the speed of convergence. Then necessary and sufficient conditions are deduced for the relative frequency to converge uniformly to the probability. These conditions do depend on the distribution properties.

The main results of the paper were stated in [1].

Let X be a set of elementary events on which a probability measure P_X is defined. Let S be a collection of random events, i.e., of subsets of the space X, which are measurable with respect to the measure P_X. Let $X^{(l)}$ denote the space of samples in X of size l. On the space $X^{(l)}$ we define a probability measure P by

$$P[Y_1 \times Y_2 \cdots \times Y_l] = P_X(Y_1)P_X(Y_2) \cdots P_X(Y_l),$$

where the Y_i are measurable subsets of X.

Each sample $x_1, \cdots, x_l$ and event $A \in S$ determines a relative frequency for A equal to the quotient of the number n_A of those elements of the sample which belongs to A and the total size l of the sample: $v_A^{(l)}(x_1, \cdots, x_l) = n_A/l$.

Bernoulli's theorem states that $|\nu_A^{(l)} - P_A| \xrightarrow{\mathbf{P}} 0$ (P_A is the probability of the event A). We shall be interested in the maximum difference over the class S between relative frequency and probability, namely,

$$\pi^{(l)} = \sup_{A \in S} |\nu_A^{(l)} - P_A|.$$

The quantity $\pi^{(l)}$ is a point function in $X^{(l)}$-space. We shall assume that this function is measurable with respect to measure in $X^{(l)}$, i.e., that $\pi^{(l)}$ is a random variable.

If the variable $\pi^{(l)}$ converges in probability to zero as the sample size l is increased indefinitely, then we shall say that the relative frequency of events $A \in S$ tends (in probability) to the probability of these events uniformly over the class S. The subsequent theorems are devoted to estimates for the probability of the events$\{\pi^{(l)} > \varepsilon\}$ and to a clarification of conditions under which, for any ε,

$$\lim_{l \to \infty} \mathbf{P}\{\pi^{(l)} > \varepsilon\} = 0.$$

1. Sufficient Conditions not Depending on Distribution Properties

1. Subsidiary definitions. Let $X_r = x_1, \cdots, x_r$ be a finite sample of elements in X. Each set A in S determines in this sample a subsample $X_r^A = x_{i_1}, \cdots, x_{i_k}$ consisting of those terms of the sample X_r which belong to A. We shall say that the set A *induces the subsample* X_r^A in the sample X_r. We denote the set of all different subsamples induced by the sets of S in the sample X_r by $S(x_1, \cdots, x_r)$ or $S(X_r)$. The number of different subsamples of the sample X_r induced by the sets in S will be termed the *index of the system S with respect to the sample* $x_1, \cdots, x_r$ and will be denoted by $\Delta^S(x_1, \cdots, x_r)$. Obviously, $\Delta^S(x_1, \cdots, x_r)$ is always at most 2^r. The function

$$m^S(r) = \max \Delta^S(x_1, \cdots, x_r),$$

where the maximum is taken over all samples of size r, will be called the *growth function.*

Example 1. Let X be a straight line and let S be the set of all rays of the form $x \leqq a$. In this case, $m^S(r) = r + 1$.

Example 2. X is the segment $[0, 1]$. S consists of all open sets. In this case, $m^S(r) = 2^r$.

Let us examine the following example which is important in the subsequent discussions.

Example 3. Let $X = E_n$, Euclidean n-space. The set S of events consists of all half-spaces of the form $(x, \varphi) \geqq 1$, where φ is a fixed vector. Let us evaluate the growth function $m^S(r)$.

Consider along with the space E_n of vectors x, the space $\bar{E}_n$ of vectors φ. To each vector $x_k \in E_n$, there corresponds a partition of the space $\bar{E}_n$ into the half-space $(x_k, \varphi) \geqq 1$ and the half-space $(x_k, \varphi) < 1$. Conversely, each vector φ determines some event in the system S.

Consider r vectors $x_1, \cdots, x_r$. They furnish a partition of $\bar{E}_n$ into a number of components such that the vectors φ inside each component determine events $A \in S$ that induce one and the same subsample in the sample $x_1, \cdots, x_r$.

Let $\Phi(n, r)$ be the maximum number of components into which it is possible to partition n-dimensional space by means of r hyperplanes.

By definition, $m^S(r) = \Phi(n, r)$. The following recurrence relation holds:

$$(1) \quad \Phi(n, r) = \Phi(n, r-1) + \Phi(n-1, r-1), \quad \Phi(0, r) = 1, \quad \Phi(n, 0) = 1.$$

In what follows essential use will be made of the function $\Phi(n, r)$.

It is not hard to show that

$$\Phi(n, r) = \begin{cases} \sum_{k=0}^{n} \binom{r}{n} & \text{if } r > n, \\ 2^r & \text{if } r \leqq n. \end{cases}$$

For $n > 0$ and $r \geqq 0$, $\Phi(n, r) \leqq r^n + 1$.

Throughout the following, we take $\binom{n}{k} = 0$ if $n < k$.

2. Properties of the growth function. The growth function for a class of events S has the following property: it is either identically equal to 2^r or is majorized by the power function $r^n + 1$, where n is a constant equaling the value of r for which the equality $m^S(r) = 2^r$ is violated for the first time. To prove this fact, we need a lemma.

Lemma 1. *If for some sample of size* $i: x_1, \cdots, x_i$ *and number* n, $1 \leqq n \leqq i$,

$$\Delta^S(x_1, \cdots, x_i) \geqq \Phi(n, i),$$

then there exists a subsample $x_{i_1}, \cdots, x_{i_n}$ *of this sample such that*

$$\Delta^S(x_{i_1}, \cdots, x_{i_n}) = 2^n.$$

$\Phi(n, i)$ *is defined by the recurrence relation* (1).

PROOF. We shall prove the lemma by induction. For $n = 1$, as well as for $n = i$, the statement of the lemma easily follows from the definition of the index $\Delta^S(x_1, \cdots, x_i)$ and the fact that, for $i \geqq 1$, $\Phi(1, i) \geqq 2$ and $\Phi(i, i) = 2^i$. Assume now that the lemma holds for all $i < r$ and $n \leqq i$ but is false for $i = r$. In other words, let there exist a sample $X_r = x_1, \cdots, x_r$ and a number $n < r$ such that

$$(2) \qquad \Delta^S(x_1, \cdots, x_r) \geqq \Phi(n, r)$$

and yet the relation $\Delta^S(x_{i_1}, \cdots, x_{i_n}) = 2^n$ does not hold for any subsample of size n. Then this relation certainly does not hold for each subsample of size n of the sample $X_{r-1} = x_1, \cdots, x_{r-1}$. But, by assumption, the lemma is valid for the sample X_{r-1} and hence

$$(3) \qquad \Delta^S(x_1, \cdots, x_{r-1}) < \Phi(n, r-1).$$

Further, all subsamples induced by the sets in S in the sample X_{r-1} may be split into two types. To the first type belongs every subsample t induced by S in X_{r-1} such that only one of the subsamples is induced in the whole sample X_r: either t or t, x_r. To the second belong those t for which both t and t, x_r are induced in the whole sample. Correspondingly, the set S is partitioned into two subsets: the subset S' which induces subsamples of the first type and the subset S'' which induces subsamples of the second type.

Let a be the number of elements in the set of subsamples of the first type and b the number of elements in the set of subsamples of the second type. Then the following relations hold:

$$\Delta^S(x_1, \cdots, x_{r-1}) = a + b, \tag{4}$$

$$\Delta^S(x_1, \cdots, x_r) = a + 2b. \tag{5}$$

Taking (3)–(5) into consideration, we have

$$\Delta^S(x_1, \cdots, x_r) < \Phi(n, r - 1) + b. \tag{6}$$

Let us now estimate the quantity $\Delta^{S''}(x_1, \cdots, x_{r-1}) = b$. To this end, observe that there exists no subsample $x_{j_1}, \cdots, x_{j_{n-1}}$ of the sample $x_1, \cdots, x_{r-1}$ for which

$$\Delta^{S''}(x_{j_1}, \cdots, x_{j_{n-1}}) = 2^{n-1}. \tag{7}$$

Equation (7) is impossible since if it were valid, so would the equation

$$\Delta^S(x_{j_1}, \cdots, x_{j_{n-1}}, x_r) = 2^n$$

be valid. The latter is impossible by virtue of the assumption made at the outset of the proof of the lemma. Thus,

$$\Delta^{S''}(x_{j_1}, \cdots, x_{j_{n-1}}) < 2^{n-1}$$

for any subsample of X_{r-1} of size $n - 1$.

But the lemma holds for the sample X_{r-1} and hence

$$b = \Delta^{S''}(x_1, \cdots, x_{r-1}) < \Phi(n - 1, r - 1). \tag{8}$$

Substituting (8) into (6), we obtain

$$\Delta^S(x_1, \cdots, x_r) < \Phi(n, r - 1) + \Phi(n - 1, r - 1).$$

Using (1), we have $\Delta^S(X_r) < \Phi(n, r)$. This inequality contradicts assumption (2). The resultant contradiction thus proves the lemma.

Theorem 1. *The growth function $m^S(r)$ is either identically equal to 2^r or else is majorized by the power function $r^n + 1$, where n is a positive constant equaling the value of r for which the equation*

$$m^S(r) = 2^r$$

is violated for the first time.

PROOF. As already mentioned, $m^S(r) \leqq 2^r$. Suppose $m^S(r)$ is not identically equal to 2^r and suppose n is the first value of r for which $m^S(r) \neq 2^r$. Then, for any sample of size $r > n$,

$$\Delta^S(x_1, \cdots, x_r) < \Phi(n, r).$$

Otherwise, on the basis of the statement of the lemma, a subsample $x_{i_1}, \cdots, x_{i_n}$ could be found such that

$$\Delta^S(x_{i_1}, \cdots, x_{i_n}) = 2^n. \tag{9}$$

But (9) is impossible, since by assumption $m^S(n) \neq 2^n$. Thus $m^S(r)$ is either identically equal to 2^r or else is majorized by $\Phi(n, r)$. In turn, for $r > 0$, $\Phi(n, r) < r^n + 1$.

3. Main lemma. Let a sample of size $2l$ be taken: $X_{2l} = \{x_1, \cdots, x_l, x_{l+1}, \cdots, x_{2l}\}$ and suppose the relative frequencies of the event $A \in S$ have been calculated in the first semi-sample $x_1, \cdots, x_l = X'_l$ and the second semi-sample $x_{l+1}, \cdots, x_{2l} = X''_l$. Let the respective frequencies be denoted by ν'_A and ν''_A and consider the difference of these quantities $\rho_A^{(l)} = |\nu'_A - \nu''_A|$. We are interested in the maximum difference between these quantities over all events in class S,

$$\rho^{(l)} = \sup_{A \in S} \rho_A^{(l)}.$$

Observe that $\sup_{A \in S} \rho_A^{(l)} = \max_{A \in S} \rho_A^{(l)}$ since, for fixed l, $\rho_A^{(l)}$ takes on only a finite number of values. Throughout the following we shall assume that $\rho^{(l)}$ is a measurable function.

In this subsection, we shall show that if $\rho^{(l)} \to 0$ as $l \to \infty$, then so does $\pi^{(l)} \xrightarrow{\mathbf{P}} 0$ and that the estimates for $\rho^{(l)}$ lead to estimates for $\pi^{(l)}$.

It is convenient to introduce the following notation:

$$Q = \{\pi^{(l)} > \varepsilon\}, \qquad C = \{\rho^{(l)} \geqq \tfrac{1}{2}\varepsilon\}.$$

Lemma 2. *For* $l > 2/\varepsilon^2$,

$$\mathbf{P}(C) \geqq \tfrac{1}{2}\mathbf{P}\{Q\}.$$

PROOF. By definition,

$$\mathbf{P}(C) = \int_{X^{(2l)}} \theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) d\mathbf{P}, \quad \text{where} \quad \theta(z) = \begin{cases} 1 & \text{for } z \geqq 0, \\ 0 & \text{for } z < 0. \end{cases}$$

Taking into account that $X^{(2l)}$ is the direct product $X'^{(l)} \times X''^{(l)}$, where $X'^{(l)}$ is the space of the first semi-samples X'_l and $X''^{(l)}$ the space of the second semi-samples X''_l, we have by Fubini's theorem that

$$\mathbf{P}(C) = \int_{X'^{(l)}} d\mathbf{P}' \int_{X''^{(l)}} \theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) d\mathbf{P}''.$$

Replacing the integration over the whole space $X'^{(l)}$ by integration over the event Q, we obtain

$$\mathbf{P}(C) \geqq \int_Q d\mathbf{P}' \int_{X''^{(l)}} \theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) d\mathbf{P}''. \tag{10}$$

By definition, to each fixed semi-sample X'_l belonging to Q, there exists an event $A_0 \in S$ such that $|P_{A_0} - \nu'_{A_0}| > \varepsilon$. Thus, to satisfy the condition $\rho^{(l)}_{A_0} \geqq \varepsilon/2$ or, equivalently, the condition $|\nu'_{A_0} - \nu''_{A_0}| \geqq \varepsilon/2$, we merely have to require that $|\nu''_{A_0} - P_{A_0}| \leqq \varepsilon/2$.

Coming back to inequality (10), we estimate the inner integral obtaining

$$\int_{X''^{(l)}} \theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) d\mathbf{P}'' \geqq \int_{X''^{(l)}} \theta\left(\rho^{(l)}_{A_0} - \frac{\varepsilon}{2}\right) d\mathbf{P}'' \geqq \int_{X''^{(l)}} \theta\left(\frac{\varepsilon}{2} - |\nu''_{A_0} - P_{A_0}|\right) d\mathbf{P}''.$$

The right-hand side of this last inequality stands for the probability that the difference between the relative frequency and the probability of a fixed event does not exceed $\frac{1}{2}\varepsilon$, i.e.,

$$\int_{X''^{(l)}} \theta\left(\frac{\varepsilon}{2} - |\nu''_{A_0} - P_{A_0}|\right) d\mathbf{P}'' = 1 - \mathbf{P}\left(|\nu''_{A_0} - P_{A_0}| > \frac{\varepsilon}{3}\right).$$

By Chebyshev's inequality applied to the binomial distribution,

$$\mathbf{P}\left(|\nu''_{A_0} - P_{A_0}| > \frac{\varepsilon}{2}\right) \leqq \frac{4(1 - P_{A_0})P_{A_0}}{\varepsilon^2 l} < \frac{1}{\varepsilon^2 l}.$$

Therefore, for $l \geqq 2/\varepsilon^2$,

$$\int_{X''^{(l)}} \theta\left(\frac{\varepsilon}{2} - |\nu''_{A_0} - P_{A_0}|\right) d\mathbf{P}'' > \frac{1}{2}.$$

From this it immediately follows that, for $l \geqq 2/\varepsilon^2$,

$$\mathbf{P}(C) \geqq 1/2\mathbf{P}(Q).$$

The lemma is proved.

4. Sufficient conditions for uniform convergence

Theorem 2. *The probability that the relative frequency of at least one event in class S differs from its probability in an experiment of size l by more then ε, for $l \geqq 2/\varepsilon^2$, satisfies the inequality*

$$\mathbf{P}(\pi^{(l)} > \varepsilon) \leqq 4m^S(2l)\, e^{-\varepsilon^2 l/8}.$$

Corollary. *A sufficient condition for the relative frequencies of events in class S to converge uniformly over S (in probability) to their corresponding probabilities is that there exist a finite n such that $m^S(l) \leqq l^n + 1$ for all l.*

Proof. By virtue of Lemma 2, it suffices to estimate

$$\mathbf{P}\left(\rho^{(l)} \geqq \frac{\varepsilon}{2}\right) = \int_{X''^{(2l)}} \theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) d\mathbf{P},$$

where $\rho^{(l)}$ is viewed as a function of the sequence

$$X_{2l} = (x_1, \cdots, x_l, x_{l+1}, \cdots, x_{2l}).$$

Consider the mapping of the space $X^{(2l)}$ onto itself resulting from some permutation T_i of the elements of the sequence X_{2l}. By virtue of the symmetry of the definition of the measure $\mathbf{P}$ on $X^{(2l)}$, the following relation holds for any integrable function $f(X_{2l})$:

$$\int_{X^{(2l)}} f(X_{2l})\, d\mathbf{P} = \int_{X^{(2l)}} f(T_i X_{2l})\, d\mathbf{P}.$$

Therefore,

$$\mathbf{P}\left\{\rho^{(l)} \geqq \frac{\varepsilon}{2}\right\} = \int_{X^{(2l)}} \frac{1}{(2l)!} \sum_i \theta\left(\rho^{(l)}(T_i X_{2l}) - \frac{\varepsilon}{2}\right) d\mathbf{P}, \tag{11}$$

where the summation is over all $(2l)!$ permutations.

Observe further that

$$\theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) = \theta\left(\sup_{A \in S} |v'_A - v''_A| - \frac{\varepsilon}{2}\right) = \sup_{A \in S} \theta\left(|v'_A - v''_A| - \frac{\varepsilon}{2}\right).$$

Clearly, if two sets A_1 and A_2 induce the same subsample in a sample $(x_1, \cdots, x_l, x_{l+1}, \cdots, x_{2l})$, then

$$v'_{A_1}(T_i X_{2l}) = v'_{A_2}(T_i X_{2l}), \qquad v''_{A_1}(T_i X_{2l}) = v''_{A_2}(T_i X_{2l})$$

and hence, $\rho^{(l)}_{A_1}(T_i X_{2l}) = \rho^{(l)}_{A_2}(T_i X_{2l})$ for any permutation T_i. This implies that if we choose the subsystem $S' \subset S$ consisting of all the sets A that induce essentially different subsamples in the sample X_{2l}, then

$$\sup_{A \in S} \theta\left(\rho^{(l)}_A(T_i X_{2l}) - \frac{\varepsilon}{2}\right) = \sup_{A \in S'} \theta\left(\rho^{(l)}_A(T_i X_{2l}) - \frac{\varepsilon}{2}\right) \leqq \sum_{A \in S'} \theta\left(\rho^{(l)}_A(T_i X_{2l}) - \frac{\varepsilon}{2}\right)$$

(the number of elements in S' is equal to $\Delta^{S'}(x_1, \cdots, x_{2l})$). These relations enable us to estimate the integrand in (11):

$$\frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}(T_i X_{2l}) - \frac{\varepsilon}{2}\right) = \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \sup_{A \in S} \theta\left(\rho^{(l)}_A(T_i X_{2l}) - \frac{\varepsilon}{2}\right)$$
$$\leqq \sum_{A \in S'} \left[\frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}_A(T_i X_{2l}) - \frac{\varepsilon}{2}\right)\right].$$

The expression in brackets denotes the quotient of the number of arrangements in a sample (of fixed composition), for which $|v'_A - v''_A| \leqq \frac{1}{2}\varepsilon$, and the overall number of permutations. It is easy to see that it is equal to

$$\Gamma = \sum_{k:\{|2k/l - m/l| \geqq \varepsilon/2\}} \frac{\binom{m}{k}\binom{2l-m}{l-k}}{\binom{2l}{l}}.$$

where m is the number of elements in the sample $x_1, \cdots, x_{2l}$ belonging to A. This expression satisfies the estimate $\Gamma \leqq 2\,e^{-\varepsilon^2 l/8}$. This estimate can be derived by a simple but long computation and so we omit the proof.

Thus,

$$\frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}(T_i X_{2l}) - \frac{\varepsilon}{2}\right) \leqq \sum_{A \in S'} 2\,e^{-\varepsilon^2 l/8} = 2\Delta^S(x_1, \cdots, x_{2l})\,e^{-\varepsilon^2 l/8}$$
$$\leqq 2m^S(2l)\,e^{-\varepsilon^2 l/8}.$$

Substituting this estimate in the integral (11), we obtain

$$\mathbf{P}\left\{\rho^{(l)} \geqq \frac{\varepsilon}{2}\right\} \leqq 2m^S(2l)\,e^{-\varepsilon^2 l/8}.$$

By virtue of Lemma 2, this yields

$$\mathbf{P}\{\pi^{(l)} > \varepsilon\} \leqq 4m^S(2l)\,e^{-\varepsilon^2 l/8}.$$

To complete the proof, it remains for us to observe that

$$m^S(2l) < (2l)^n + 1, \qquad \lim_{l\to\infty} \mathbf{P}\{\pi^{(l)} > \varepsilon\} \leqq 4 \lim_{l\to\infty} [1 + (2l)^n]\,e^{-\varepsilon^2 l/8} = 0.$$

The resultant sufficient condition does not depend on the distribution properties.

5. On uniform convergence with probability one and estimation of the sample size. In the preceding subsection, we gave sufficient conditions for the relative frequencies to converge uniformly over the class of events S to the probabilities. In this subsection, we shall show that the resultant conditions assure uniform convergence almost surely. In proving this, we make use of the following well-known lemma of probability theory (cf. [2]):

If for any positive ε

$$\sum_i^\infty \mathbf{P}(|\xi_i - C| > \varepsilon) < \infty,$$

then

$$\mathbf{P}(\xi_i \to C) = 1.$$

Theorem 3. *If* $m^S(l) \leqq l^n + 1$, *then* $\mathbf{P}(\pi^{(l)} \to 0) = 1$.

Proof. Since

$$\mathbf{P}(\pi^{(l)} > \varepsilon) \leqq 4m^S(2l)\,e^{-\varepsilon^2 l/8}$$

for $l > l^* = 2/\varepsilon^2$, the series

$$\sum_{l=1}^{\infty} \mathbf{P}(\pi^{(l)} > \varepsilon) \leqq \sum_{l=1}^{l^*} \mathbf{P}(\pi^{(l)} > \varepsilon) + 4 \sum_{l=l^*+1}^{\infty} [(2l)^n + 1]e^{-\varepsilon^2 l/8}$$

is convergent for any ε. By the lemma, this implies that

$$\mathbf{P}(\pi^{(l)} \to 0) = 1.$$

Example (Glivenko's theorem). As in Example 1 of Subsection 1, let X be the real line, $-\infty < x < \infty$. The set S is given by all rays of the form $x \leqq a$.

As was shown, in this case $m^S(l) = l + 1$ and hence uniform convergence holds with probability one. Set

$$A = \{x \leqq a\}, \qquad \mathbf{P}_A = F(a); \qquad v_A^n = F_n^{(a)}.$$

In this notation, the fact that there is uniform convergence with probability one may be written in the form

$$\mathbf{P}(\sup_a |F_n(a) - F(a)| \to 0) = 1.$$

This formula makes up the content of Glivenko's theorem (cf. [2]).

In a similar way, we can satisfy ourselves that uniform convergence with probability one also holds for the class of events considered in Example 3 of Subsection 1.

The class of events considered in Example 2 does not satisfy the sufficient conditions.

In many applications, it is important to know what the sample size must be in order that, with probability at least $(1 - \eta)$, one could assert that the relative frequencies differ from their corresponding probabilities by an amount less than ε simultaneously over the entire class of events.

In other words, beginning with what value l, does the following inequality hold:

$$4m^S(2l)\, e^{-\varepsilon^2 l/8} \leqq \eta \quad \text{if } m^S(l) \leqq l^n + 1?$$

It is possible to show that this inequality holds when

$$l \geqq \frac{16}{\varepsilon^2}\left(n \log \frac{16n}{\varepsilon^2} - \log \frac{\eta}{4}\right).$$

2. Necessary and Sufficient Conditions

6. Some additional properties of the index. Observe first that the definition of an index immediately implies that

$$(12) \quad \Delta^S(x_1, \cdots, x_k, x_{k+1}, \cdots, x_l) \leqq \Delta^S(x_1, \cdots, x_k)\Delta^S(x_{k+1}, \cdots, x_l).$$

Hence it follows that

$$\log_2 \Delta^S(x_1, \cdots, x_k, x_{k+1}, \cdots, x_l) \leqq \log_2 \Delta^S(x_1, \cdots, x_k) + \log_2 \Delta^S(x_{k+1}, \cdots, x_l). \tag{13}$$

In what follows it will be assumed that the index $\Delta^S(x_1, \cdots, x_l)$ viewed as a function of $X_l = \{x_1, \cdots, x_l\}$ is measurable with respect to the measure $\mathbf{P}$.

Let

$$F^{(l)}(z) = \mathbf{P}(\log_2 \Delta^S(x_1, \cdots, x_l) < z), \qquad \mathbf{E} \log_2 \Delta^S(x_1, \cdots, x_l) = H^S(l).$$

$H^S(l)$ is the entropy of the system of events S in samples of size l.

Inequality (13) implies that

$$H^S(l_1 + l_2) \leqq H^S(l_1) + H^S(l_2).$$

The following lemma is proved in the same way as in [3].

Lemma 3. *The sequence $H^S(l)/l$ has a limit c, $0 \leqq c \leqq 1$, as $l \to \infty$.*

Let us now show that for large l the distribution of the random variable $\xi^{(l)} = l^{-1} \log_2 \Delta^S(x_1, \cdots, x_l)$ is concentrated near c.

Lemma 4. $\lim_{l \to \infty} \mathbf{P}(|\xi^{(l)} - c| > \varepsilon) = 0$ *for* $\varepsilon > 0$.

Proof. Denote $\mathbf{P}(|\xi^{(l)} - c| > \varepsilon)$ by $\mathbf{P}(l, \varepsilon)$, $\mathbf{P}(\xi^{(l)} - c > \varepsilon)$ by $P^+(l, \varepsilon)$ and $\mathbf{P}(c - \xi^{(l)} > \varepsilon)$ by $P^-(l, \varepsilon)$. Accordingly,

$$P(l, \varepsilon) = P^+(l, \varepsilon) + P^-(l, \varepsilon).$$

Lemma 3 implies the existence of an l_0 such that

$$\left| \frac{H^S(l_0)}{l_0} - c \right| < \frac{\varepsilon}{4}. \tag{14}$$

We first estimate $P^+(l, \frac{1}{2}\varepsilon)$ with $l = nl_0$ (n an integer).

From (13) it follows that

$$\log_2 \Delta^S(x_1, \cdots, x_{nl_0}) \leqq \sum_{i=0}^{n-1} \log_2 \Delta^S(x_{il_0+1}, \cdots, x_{(i+1)l_0}).$$

From this we obtain

$$P^+(nl_0, \tfrac{1}{2}\varepsilon) \leqq \mathbf{P}\left\{\frac{1}{nl_0} \sum_{i=0}^{n-1} \log_2 \Delta^S(x_{il_0+1}, \cdots, x_{(i+1)l_0}) - c > \frac{\varepsilon}{2}\right\}. \tag{15}$$

Let

$$y = \frac{1}{nl_0} \sum_{i=0}^{n-1} \log_2 \Delta^S(x_{il_0+1}, \cdots, x_{(i+1)l_0})$$

and observe that

$$\sum_{i=0}^{n-1} \log_2 \Delta^S(x_{il_0+1}, \cdots, x_{(i+1)l_0})$$

is a sum of independent random variables with expectation $H^S(l_0)$ and a certain variance D. Hence it follows that

$$\mathbf{E}y = \frac{H^S(l_0)}{l_0}; \qquad \mathbf{D}y = \frac{D}{nl_0^2}.$$

Using inequality (14) and Chebyshev's inequality, we obtain

$$\mathbf{P}\left(y - c > \frac{\varepsilon}{2}\right) \leqq \mathbf{P}\left(y - \frac{H^S(l_0)}{l_0} > \frac{\varepsilon}{4}\right) \leqq \mathbf{P}\left(|y - My| > \frac{\varepsilon}{4}\right) \leqq \frac{16D}{n\varepsilon^2 l_0^2}.$$

This with the help of (15) leads to

$$P^+\left(nl_0, \frac{\varepsilon}{2}\right) \leqq \frac{16D}{n\varepsilon^2 l_0^2} \quad \text{and} \quad \lim_{n\to\infty} P^+\left(nl_0, \frac{\varepsilon}{2}\right) = 0.$$

Let us now prove that

$$\lim_{l\to\infty} P^+(l, \varepsilon) = 0.$$

For arbitrary $l > l_0$, let n be such that $nl_0 < l < (n + 1)l_0$. We have

$$\frac{1}{nl_0}\log_2 \Delta^S(x_1, \cdots, x_{(n+1)l_0}) > \frac{1}{l}\log_2 \Delta^S(x_1, \cdots, x_l).$$

This leads to

$$\mathbf{P}\left(\frac{n+1}{nl_0}\xi^{(n+1)l_0} > c + \varepsilon\right) > P^+(l, \varepsilon).$$

But, for sufficiently large n,

$$\mathbf{P}\left(\frac{n+1}{n}\xi^{(n+1)l_0} > c + \varepsilon\right) \leqq \mathbf{P}\left(\xi^{(n+1)l_0} > c + \frac{\varepsilon}{2}\right) = P^+\left((n + 1)l_0, \frac{\varepsilon}{2}\right).$$

Therefore,

$$\lim_{l\to\infty} P^+(l, \varepsilon) = 0. \tag{16}$$

We next prove that $P^-(l, \varepsilon) \to 0$ as $l \to \infty$.

From the properties of expectation and the fact that $\mathbf{E}\xi^{(l)} = H^S(l)/l$, it follows that

$$\int_0^{H^S(l)/l} \left(\frac{H^S(l)}{l} - \xi\right) dF_\xi = \int_{H^S(l)/l}^1 \left(\xi - \frac{H^S(l)}{l}\right) dF_\xi. \tag{17}$$

Denoting the right-hand side of (17) by R_2 and the left-hand side by R_1, we estimate them assuming that l is so large that $|H^S(l)/l - c| < \varepsilon/2$ and obtain first

$$R_1 \geqq \frac{\varepsilon}{2}\int_0^{c-\varepsilon} dF_\xi = \frac{\varepsilon}{2}P^-(l, \varepsilon). \tag{18}$$

Let δ be a positive number. Then

$$\begin{aligned}R_2 &\leqq \left|\int_{H^S(l)/l}^{c+\delta}\left(\xi - \frac{H^S(l)}{l}\right)dF_\xi\right| + \int_{c+\delta}^{1}\left(\xi - \frac{H^S(l)}{l}\right)dF_\xi \\ &\leqq \left|c + \delta - \frac{H^S(l)}{l}\right| + P^+(l, \delta).\end{aligned} \tag{19}$$

Combining the estimates (18) and (19), we have

$$P^-(l, \varepsilon) \leqq \frac{2}{\varepsilon}\left[\left|c + \delta - \frac{H^S(l)}{l}\right| + P^+(l, \delta)\right].$$

This in conjunction with Lemma 3 and (16) implies that

$$\lim_{l\to\infty} P^-(l, \varepsilon) \leqq \frac{2\delta}{\varepsilon}$$

and since δ is arbitrary, that

$$\lim_{l\to\infty} P^-(l, \varepsilon) = 0. \tag{20}$$

Finally, according to (16) and (20),

$$\lim_{l\to\infty} P(l, \varepsilon) = 0.$$

The lemma is proved.

7. Necessary and sufficient conditions

Theorem 4. *A necessary and sufficient condition for the relative frequencies to converge* (*in probability*) *to the probabilities uniformly over the class of events S is that*

$$\lim_{l\to\infty} \frac{H^S(l)}{l} = 0. \tag{21}$$

Observe that, by Lemma 4, condition (21) is equivalent to the fact that

$$\lim_{l\to\infty} \mathbf{P}\left(\frac{1}{l}\log_2 \Delta^S(x_1, \cdots, x_l) > \delta\right) = 0 \tag{22}$$

for all $\delta > 0$.

Proof of sufficiency. Suppose

$$\lim_{l\to\infty} \frac{H^S(l)}{l} = 0.$$

It will be recalled that, by the lemma, $2\mathbf{P}(C) \geqq \frac{1}{2}P(Q)$. Let us estimate the probability of event C.

As we showed in Subsection 4,

$$\mathbf{P}(C) \leqq \frac{1}{(2l)!}\int_{X^{(2l)}}^{4} \sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}(T_iX_{2l}) - \frac{\varepsilon}{2}\right) d\mathbf{P}.$$

Let $\delta = \varepsilon^2/16$ and split the region of integration into two parts: $X_1^{(2l)} = \{\log_2 \Delta^S(X_{2l}) \leqq 2\delta\}$ and $X_2^{(2l)} = X^{(2l)} - X_1^{(2l)}$. Then

$$\mathbf{P}(C) = \int_{X_1^{(2l)}} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}(T_iX_{2l}) - \frac{\varepsilon}{2}\right) d\mathbf{P}$$
$$+ \int_{X_2^{(2l)}} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}(T_iX_{2l}) - \frac{\varepsilon}{2}\right) d\mathbf{P}.$$

Since the integrand does not exceed unity, we have

$$\mathbf{P}(C) \leqq \int_{X_1^{(2l)}} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}(T_iX_{2l}) - \frac{\varepsilon}{2}\right) d\mathbf{P} + P^+(2l, \delta).$$

In Subsection 4 it was shown that

$$\frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}(T_iX_{2l}) - \frac{\varepsilon}{2}\right) \leqq 2\Delta^S(x_1, \cdots, x_{2l})\, e^{-\varepsilon^2 l/8}.$$

Using the fact that $\Delta^S(x_1, \cdots, x_{2l}) \leqq 2^{2\delta l}$ in the region of integration, we have

$$\mathbf{P}(C) \leqq 2 \cdot 2^{2\delta l} e^{-\varepsilon^2 l/8} + P^+(2l, \delta) = 2(2/e)^{+\varepsilon^2 l/8} + P^+(2l, \delta).$$

But, by Lemma 4, $\lim_{l\to\infty} P^+(2l, \delta) = 0$. Hence it follows that $\lim_{l\to\infty} \mathbf{P}(C) = 0$ and so $\lim_{l\to\infty} \mathbf{P}(Q) = 0$. The sufficiency is proved.

Proof of necessity. 1°. Suppose

$$\lim_{l\to\infty} \frac{H^S(l)}{l} = c > 0. \tag{23}$$

To prove the necessity, we must show that there exists a positive ε such that

$$\lim_{l\to\infty} \mathbf{P}(Q) = \lim_{l\to\infty} \mathbf{P}\{\sup_{A\in S}|\nu_A^{(l)} - P_A| > \varepsilon\} \neq 0.$$

It suffices to estimate the probability of the event

$$C' = \{\sup|\nu'_A - \nu''_A| > 2\varepsilon\}.$$

Indeed, we shall show that from a lower estimate for the probability of event C' will follow a lower estimate for $\mathbf{P}(Q)$. Suppose that $x_1, \cdots, x_{2l}$ is a given sample and that the event Q does not occur on both semi-samples, i.e.,

$$\sup_{A\in S}|\nu'_A - P_A| \leqq \varepsilon, \qquad \sup_{A\in S}|\nu''_A - P_A| \leqq \varepsilon.$$

Then automatically $\sup_{A\in S}|\nu'_A - \nu''_A| \leqq 2\varepsilon$. Thus, taking into account the independence of the semi-samples, we obtain

$$1 - \mathbf{P}(C') \geqq (1 - \mathbf{P}(Q))^2, \text{i.e.}, \mathbf{P}(C') \leqq 2\mathbf{P}(Q) - \mathbf{P}^2(Q).$$

A weakening of this inequality yields $\mathbf{P}(Q) \geqq \frac{1}{2}\mathbf{P}(C')$.

2°. Observe now that, by virtue of Lemma 1, one can find a subsample $x_{i_1}, \cdots, x_{i_n}$ of X_{2l} such that S induces in it all possible subsamples providing

$$\Delta^S(x_1, \cdots, x_1) \geqq \Phi(n, l). \tag{24}$$

We assign some q, $0 < q < \frac{1}{4}$, and we estimate the probability of (24) holding for $n = [ql]$. It is not hard to see that, for $q < \frac{1}{4}$ and $n = [ql]$,

$$\Phi(n, l) = \sum_{i=0}^{n} \binom{l}{i} < 2\binom{l}{n} < 2\frac{l^{[ql]}}{[ql]!}.$$

In what follows, we shall assume that $l \geqq 1/q$. Thus $[ql] \geqq \frac{1}{2}ql$. Applying Stirling's formula, we obtain the estimate

$$\Phi(n, l) < 2\left(\frac{2e}{q}\right)^{ql}.$$

Now for the probability that (24) holds, we obtain the estimate

$$\mathbf{P}\{\Delta^S(x_1, \cdots, x_l) \geqq \Phi(n, l)\} \geqq \mathbf{P}\left\{\Delta^S(X_l) > \left(\frac{2e}{q}\right)^{ql}\right\}$$
$$= \mathbf{P}\left\{\frac{\log_2 \Delta^S(x_1, \cdots, x_l)}{l} > q\log_2\frac{2e}{q} + \frac{1}{l}\right\}.$$

Since $\lim_{l\to\infty} H^S(l)/l = c$, we can choose a sufficiently small positive q such that

$$q\log_2\frac{2e}{q} < c. \tag{25}$$

Assuming further that (25) is satisfied, we can apply Lemma 4 to obtain

$$\lim_{l\to\infty} \mathbf{P}\{\Delta^S(x_1, \cdots, x_l) > \Phi(n, l)\} = 1. \tag{26}$$

3°. To complete the proof of the necessity, we just have to estimate

$$\mathbf{P}(C') = \int_{X^{(2l)}} \theta(\sup_{A\in S}|v'_A - v''_A| - 2\varepsilon)\, d\mathbf{P} = \int_{X^{(2l)}} \frac{1}{(2l)!} \sum_{i=1} \theta(\rho^l(T_iX_{2l}) - 2\varepsilon)\, d\mathbf{P}$$

for $\varepsilon > 0$.

Choose a q satisfying (25) and let B denote the set of those samples for which $\Delta^S(x_1, \cdots, x_{2l}) \geqq \Phi([2ql], 2l)$. Then

$$\mathbf{P}(C') \geqq \int_B \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta(\rho^{(l)}(T_iX_{2l}) - 2\varepsilon)\, d\mathbf{P} = \int_B Z\, d\mathbf{P}.$$

Let us examine the integrand Z assuming that $X_{2l} \in B$.

Observe that all permutations T_i can be classified into groups R_i corresponding to the same partition into the first and second semi-sample. The value of $\rho^{(l)}(T_i X_{2l})$ does not change within the framework of one group. The number of permutations in all the groups is the same and equal to $(l!)^2$. The number of groups is $\binom{2l}{l}$. Thus,

$$Z = \frac{1}{\binom{2l}{l}} \sum_{i=1}^{\binom{2l}{l}} \theta(\rho^{(l)}(R_i X_{2l}) - 2\varepsilon).$$

By Lemma 1, taking into consideration that X_{2l} satisfies (24) we can pick out a subsample y in this sample of size n such that S induces all possible subsamples in it. The partition R_i is completely prescribed if the partition N_k of the subsample y and the partition M_j of the subsample $X_{2l} - y$ are given.

Let $R_i = N_k M_j$. Let $r(k)$ be the number of elements in the subsample y which belong, under the partition N_k, to the first semi-sample and $s(j)$ the number of elements of subsample $X_{2l} - y$ which belong, under partition M_j, to the first semi-sample. Clearly, $r(k) + s(j) = l$ for k and j corresponding to the same partition R_i. We have

$$Z = \frac{1}{\binom{2l}{l}} \sum_k \sum_j{}' \theta(\rho^{(l)}(N_k M_j X_{2l}) - 2\varepsilon),$$

where $\sum_j'$ is summation over just those j for which $S(j) = l - r(k)$, and

$$Z = \frac{1}{\binom{2l}{l}} \sum_{r=0}^{l} \left(\sum_k{}' \sum_j{}' \theta(\rho^{(l)}(N_k M_j X_{2l}) - 2\varepsilon \right),$$

where $\sum_k'$ is summation over just those k for which $r(k) = r$. For each N_k, we can specify a set $A(k) \in S$ such that $A(k)$ includes exactly the elements of subsample y which belong under partition N_k to the first semi-sample.

Introduce the notation: $t(k)$ is the number of elements in subsample $X_{2l} - y$ belonging to $A(k)$, $u(k,j)$ is the number of elements in $X_{2l} - y$ in $A(k)$ belonging, under partition M_j, to the first semi-sample. Then $v'_{A(k)} = (r + u)/l$ and $v''_{A(k)} = (t - u)/l$. Correspondingly,

$$\rho^{l}_{A(k)} = |v'_{A(k)} - v''_{A(k)}| = l^{-1}|2u + r - t|.$$

We further take into account that $\sup_{A \in S} \rho_A \geqq \rho_{A(k)}$ and replacing $\sup_{A \in S} \rho_A$ by $\rho_{A(k)}$ we estimate Z to obtain

$$Z \geqq \frac{1}{\binom{2l}{l}} \sum_{r=0}^{l} \sum_k{}' \left(\sum_j{}' \theta(l^{-1}(2u(k,j) + r - t(k)) - 2\varepsilon) \right).$$

Observe that the number of partitions N_j satisfying the condition $S(j) = l - r$ for fixed r is $\binom{2l-[2ql]}{l-r}$ and the number of partitions N_j which in addition correspond to the same u for fixed r and $A(k)$ is

$$\binom{t(k)}{u} \binom{2l - [2ql] - t(k)}{l - r - u}.$$

Using these relations, we obtain

$$Z \geqq \frac{1}{\binom{2l}{l}} \sum_{r=0}^{l} \binom{2l-[2ql]}{l-r} \sum_{k}' \sum_{u}' \frac{\binom{t(k)}{u}\binom{2l-[2ql]-t(k)}{l-r-u}}{\binom{2l-[2ql]}{l-r}},$$

where $\sum_u'$ is summation over just those u for which $l^{-1}|2u + r + t(k)| > 2\varepsilon$. The expression in the last sum is nothing else than the probability of drawing u black balls from an urn containing $2l - [2ql]$ balls of which t are black, assuming that $l - r$ balls altogether are drawn without replacement. Moreover (cf. [4]),

$$\mathbf{E}u = \frac{l - r}{2l - [2ql]}t; \qquad \mathbf{D}u \leqq l.$$

Now applying Chebyshev's inequality, we obtain

$$\mathbf{P}\left(\left|\frac{M(u) - u}{l}\right| \leqq \varepsilon\right) \geqq 1 - \frac{1}{l\varepsilon^2}$$

or

$$\sum_{u}'' \frac{\binom{t}{u}\binom{2l-[2ql]-t}{l-r-u}}{\binom{2l-[2ql]}{l-r}} \geqq 1 - \frac{1}{\varepsilon^2 l},$$

where the summation is over all u satisfying

$$\left|u - \frac{(l - r)t}{2l - [2ql]}\right| \leqq \varepsilon l. \tag{27}$$

By direct verification it is easy to show that, for $7\varepsilon \leqq r/l \leqq q + \varepsilon$ and $l > 1/\varepsilon$, inequality (27) implies that $|2u + r - t| > 2\varepsilon l$ for all t, $0 \leqq t \leqq 2l - [2ql]$. Thus, under these conditions,

$$\sum_{u}' \frac{\binom{t}{u}\binom{2l-[2ql]-t}{l-u-r}}{\binom{2l-[2ql]}{l-r}} \geqq 1 - \frac{1}{l\varepsilon^2}.$$

Coming back to the estimation of Z, we obtain for $l > 1/\varepsilon$

$$Z \geqq \frac{1}{\binom{2l}{l}} \sum_{7\varepsilon \leqq r/l \leqq q+\varepsilon} \binom{2l-[2ql]}{l-r} \sum_{k}' \left(1 - \frac{1}{l\varepsilon^2}\right)$$

$$= \frac{(1 - 1/l\varepsilon^2)}{\binom{2l}{l}} \sum_{7\varepsilon \leqq r/l \leqq q+\varepsilon} \binom{2l-[2ql]}{l-r}\binom{[2ql]}{r}.$$

Observe that

$$\lim_{l\to\infty} \frac{1}{\binom{2l}{l}} \sum_{7\varepsilon \leqq r/l \leqq q+\varepsilon} \binom{2l-[2ql]}{l-r}\binom{[2ql]}{r} = 1$$

(see, for example, the estimation of Γ in Subsection 4) if

$$0 < \varepsilon < q/7. \tag{28}$$

Finally, assuming that (28) holds, we have for $l > 1/\varepsilon$

$$\mathbf{P}(C') = \int_B Z\, d\mathbf{P} \geqq \left(1 - \frac{1}{\varepsilon^2 l}\right) \frac{\binom{2l-[2ql]}{l-r}\binom{[2ql]}{r}}{\binom{2l}{l}} \mathbf{P}(B)$$

and

$$\lim_{l\to\infty} \mathbf{P}(C') \geqq \lim_{l\to\infty} \mathbf{P}(B) = \lim_{l\to\infty} P(\Delta^S(x_1, \cdots, x_{2l}) > \Phi([2ql]2l)).$$

We showed in 2° that this last limit has the value 1. Hence it follows that $\lim_{l\to\infty} \mathbf{P}(C') = 1$. According to 1°, this then means that

$$\lim_{l\to\infty} \mathbf{P}\{\sup_{A\in S} |\nu_A^l - P_A| > \varepsilon\} = 1, \tag{29}$$

providing

$$\varepsilon < \frac{q}{7} \quad \text{and} \quad q \log_2 \frac{2e}{q} < c.$$

Thus, it is possible to choose a positive ε so that (29) holds. The theorem is proved.

Received by the editors
May 7, 1969

REFERENCES

[1] V. N. Vapnik and A. Ya. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Dokl. Akad. Nauk SSSR, 181, 4 (1968), p. 781. (In Russian.)

[2] B. V. Gnedenko, *Theory of Probability*, Chelsea Publishing Co., N.Y., 4th ed., 1968.

[3] A. Ya. Khinchin, *On basic theorems of information theory*, Uspekhi Mat. Nauk, XI, 1 (1956), pp. 17–75. (In Russian.)

[4] W. Feller, *Introduction to Probability Theory and Its Applications*, Vol. 1, Wiley and Sons, N.Y. 2nd ed., 1957.