ON PROBABILITIES OF LARGE DEVIATIONS

WASSILY HOEFFDING
UNIVERSITY OF NORTH CAROLINA

1. Summary

The paper is concerned with the estimation of the probability that the empirical distribution of n independent, identically distributed random vectors is contained in a given set of distributions. Sections 1–3 are a survey of some of the literature on the subject. In section 4 the special case of multinomial distributions is considered and certain results on the precise order of magnitude of the probabilities in question are obtained.

2. The general problem

Let $X_1, X_2, \cdots$ be a sequence of independent m-dimensional random vectors with common distribution function (d.f.) F. If we want to obtain general results on the behavior of the probability that $X^{(n)} = (X_1, \cdots, X_n)$ is contained in a set A^* when n is large, we must impose some restrictions on the class of sets. One interesting class consists of the sets A^* which are symmetric in the sense that if $X^{(n)}$ is in A^*, then every permutation $(X_{j_1}, \cdots, X_{j_n})$ of the n component vectors of $X^{(n)}$ is in A^*. The restriction to symmetric sets can be motivated by the fact that under our assumption all permutations of $X^{(n)}$ have the same distribution. Let $F_n = F_n(\cdot|X^{(n)})$ denote the empirical d.f. of $X^{(n)}$. The empirical distribution is invariant under permutations of $X^{(n)}$, and for any symmetric set A^* there is at least one set A in the space $\mathcal{G}$ of m-dimensional d.f.'s such that the events $X^{(n)} \in A^*$ and $F_n(\cdot|X^{(n)}) \in A$ are equivalent. The latter event will be denoted by $F_n \in A$ for short. Thus when we restrict ourselves to symmetric sets, we may as well consider the probabilities $P\{F_n \in A\}$, where $A = A_n$ may depend on n. (It is understood that $A \subset \mathcal{G}$ is such that the set $\{x^{(n)}|F_n(\cdot|x^{(n)}) \in A\}$ is measurable.) Since F_n converges to F in a well-known sense (Glivenko-Cantelli theorem), we may say that $P\{F_n \in A_n\}$ is the probability of a large deviation of F_n from F if F is not in A_n and not "close" to A_n, implying that $P\{F_n \in A_n\}$ approaches 0 as $n \to \infty$. For certain classes of sets A_n estimates of $P\{F_n \in A_n\}$

This research was supported in part by the Mathematics Division of the Air Force Office of Scientific Research. Part of the work was done while the author was a visiting professor at the Research and Training School of the Indian Statistical Institute in Calcutta under the United Nations Technical Assistance Program.