The Annals of Statistics
1974, Vol. 2, No. 2, 362-366

LARGE DEVIATIONS OF EMPIRICAL PROBABILITY MEASURES

BY M. STONE

University College London

Sanov's statement of first-order asymptotic behaviour of probabilities of large deviations of an empirical distribution function is here established for empirical probability measures, with attendant simplification of conditions. For the case of distribution functions, our theorem is strictly more general than a specialisation of results of Hoadley.

1. Introduction and summary. Sanov (1957) stated the following theorem: *Let F_n be the empirical distribution function for n independent observations with distribution function F. If Ω is an F-distinguishable class of distribution functions Φ then*

$$P(F_n \in \Omega) = \exp\{-nI(\Omega) + o(n)\}$$

where $I(\Omega) =_{\text{def}} \inf_{\Phi \in \Omega} \int d\Phi \log(d\Phi/dF)$.

The condition of "F-distinguishability" is too technical to describe here; Sanov suggests that it is not too difficult to see in applications whether it is satisfied or not.

For the present generalisation, we have n independent observations $x_1, \cdots, x_n$ of a general random variable x distributed with probability distribution P on a space X. The *empirical probability measure* P_n results from the allocation of measure $1/n$ to each of the points $x_1, \cdots, x_n$. We will be interested in the first-order asymptotic behaviour, as $n \rightarrow \infty$, of $P(P_n \in \Omega)$ where Ω is a fixed subset of the set $\mathscr{P} = \{Q\}$ of all probability measures Q on X. Our main result will be of interest only in the case when P is outside Ω in the sense that $I(\Omega) =_{\text{def}} \inf_{Q \in \Omega} I(Q)$, where $I(Q) =_{\text{def}} \int dQ \log(dQ/dP)$ is such that $I(\Omega) > 0$, but P is not so far outside that $I(\Omega) = \infty$. This is the case in which the event $P_n \in \Omega$ with n large represents a "large deviation" of P_n.

We will assume, without statement of sufficient conditions, that Ω is reasonable enough for $P(P_n \in \Omega)$ to be defined for all n. If it were not, we would not be interested in it.

Additional notation. D_k will denote a k-class partition $X_1^k \cup \cdots \cup X_k^k$ of X such that

$$P(X_i^k) > 0\,, \quad i = 1, \cdots, k\,;$$

$$I(Q, D_k) =_{\text{def}} \sum_i Q(X_i^k) \log\{Q(X_i^k)/P(X_i^k)\}\,;$$

$$I(\Omega, D_k) =_{\text{def}} \inf_{Q \in \Omega} I(Q, D_k)\,;$$

$$I(\Omega, k) =_{\text{def}} \sup_{D_k} \{I(\Omega, D_k) \mid k\}\,;$$

$$I(\Omega, \sup) =_{\text{def}} \sup_k I(\Omega, k)\,.$$

Received August 1972; revised May 1973.

AMS 1970 *subject classifications.* Primary 62.15; Secondary 60.30.

Key words and phrases. Large deviations, empirical probability measures, F-distinguishability, Cramér condition.

Our condition, to replace Sanov's "F-distinguishability", is (C) = (C1) ∪ (C2) ∪ (C3) given by:

(C1) $I(\Omega) < \infty$.

For arbitrary $\varepsilon > 0$, there is a probability distribution R in Ω, an integer h, a partition D_h and $\delta > 0$ such that

(C2) $I(\Omega, D_h) \leqq I(R, D_h) < I(\Omega, D_h) + \varepsilon$,
(C3) $\{Q \mid \max_i |Q(X_i^h) - R(X_i^h)| < \delta\} \subset \Omega$.

In Section 2 we prove:

THEOREM 1. *Under* (C), $P(P_n \in \Omega) = \exp\{-nI(\Omega) + o(n)\}$.

The proofs invoke Hoeffding's (1965) "slight elaboration" of one of Sanov's results for multinomial distributions.

For the case of distribution functions, the following sufficient condition for (C) is established in Section 3:

THEOREM 2. *If* (i) $X = R^1$, $P \sim F_0$, *a continuous distribution function, and* $I(\Omega) < \infty$ (ii) *there is a real-valued function* $T(F)$, *defined on* $\mathscr{F}$, *that is uniformly continuous with respect to the metric* $s(F, G) = \sup_x |F(x) - G(x)|$, *with* $\Omega = \{F \mid T(F) \geqq 0\}$, (iii) $I_r =_{\mathrm{def}} I(\Omega_r)$, *where* $\Omega_r =_{\mathrm{def}} \{F \mid T(F) \geqq r\}$, *is continuous at* $r = 0$ *then* (C) *obtains.*

(The obvious R^p version of Theorem 2 is obtainable by minor changes in the proof.)

2. Proof of Theorem 1. We note the following properties of the I functions:

(P1) $I(Q, D_k) \leqq I(Q, D_{k'}) \leqq I(Q)$ and $I(\Omega, D_k) \leqq I(\Omega, D_{k'}) \leqq I(\Omega)$ if $D_{k'}$ is a refinement of D_k;
(P2) $I(\Omega, k)$ is a non-decreasing function of k;
(P3) $I(\Omega, \sup) \leqq I(\Omega)$.

LEMMA 2.1. *Without condition*

$$P(P_n \in \Omega) \leqq \exp\{-nI(\Omega, k) + O(\log n)\} \tag{2.1}$$

where O depends only on k.

PROOF. Fix k. Then, for every D_k,

$$P_n \in \Omega \Rightarrow I(P_n, D_k) \geqq I(\Omega, D_k)\,. \tag{2.2}$$

The right-hand side of (2.2) is a condition on $(P_n(X_i^k);\ i = 1, \cdots, k)$, the empirical multinomial proportion vector defined by D_k. Use of Hoeffding's (1965) Theorem 2.1 then gives $P(P_n \in \Omega) \leqq \exp\{-nI(\Omega, D_k) + O(\log n)\}$ where O depends only on k, whence the result.

COROLLARY 2.1. $P(P_n \in \Omega) \leqq \exp\{-nI(\Omega, \sup) + o(n)\}$.

LEMMA 2.2. *Under* (C), $P(P_n \in \Omega) \geqq \exp\{-nI(\Omega, \sup) + o(n)\}$.

PROOF. Fix $\varepsilon > 0$ and obtain R, h, D_h and δ from (C). There clearly exists an integer m such that, for each $n \geqq m$, an n-point empirical probability measure S_n can be found with

$$\max_i |S_n(X_i^h) - R(X_i^h)| < \delta . \tag{2.3}$$

Take $n \geqq m$. Let $\mathscr{S}_n = \{n\text{-point empirical probability measures } Q_n \text{ such that } Q_n(X_i^h) = S_n(X_i^h),\ i = 1, \cdots, h\}$. Note that, by (C3), $\mathscr{S}_n \subset \Omega$. For $Q_n \in \mathscr{S}_n$, we have $I(Q_n, D_h) = I(S_n, D_h)$ and we find

$$|I(S_n, D_h) - I(R, D_h)| < \delta \textstyle\sum_i |\log P(X_i^h)| \\ + h \max \{|\delta \log \delta|, |(1 - \delta) \log (1 - \delta)|\} .$$

We may take δ so small that

$$|I(S_n, D_h) - I(R, D_h)| < \varepsilon . \tag{2.4}$$

Now $\mathscr{S}_n$ corresponds to a singleton A in Hoeffding's equation (2.9) with our n replacing Hoeffding's N, h replacing Hoeffding's k and $(P(X_i^h);\ i = 1, \cdots, h)$ replacing p. Whence $P(P_n \in \Omega) \geqq P(P_n \in \mathscr{S}_n) = \exp\{-nI(S_n, D_h) + O(\log n)\}$ where O depends only on h. (The uniformity of the O in Hoeffding's (2.9) with respect to A allows dependence of $\mathscr{S}_n$ on n.) Therefore, by (C2) and (2.4),

$$P(P_n \in \Omega) \geqq \exp[-n\{I(\Omega, D_h) + 2\varepsilon\} + O(\log n)] \\ \geqq \exp[-n\{I(\Omega, \sup) + 2\varepsilon\} + O(\log n)]$$

where O depends only on h which is determined by choice of ε. For fixed ε, we take n so large that $|O(\log n)| < \varepsilon n$ and the result follows.

LEMMA 2.3. *Under* (C), $I(\Omega, \sup) = I(\Omega)$.

PROOF. If not, then, by (P3), $I(\Omega, \sup) < I(\Omega)$. Let $\varepsilon = \frac{1}{2}[I(\Omega) - I(\Omega, \sup)]$. With this ε, obtain R, h, D_h from (C). Define $\bar{R} \in \mathscr{P}$ by $\bar{R}(X_i^h) = R(X_i^h)$ and $\bar{R}(B_i) = R(X_i^h)P(B_i)/P(X_i^h)$ for $B_i \subset X_i^h$, $i = 1, \cdots, h$. Then, by (C3), $\bar{R} \in \Omega$. Moreover, $I(\bar{R}) = I(\bar{R}, D_h) = I(R, D_h) < I(\Omega, D_h) + \varepsilon$ by (C2). So $I(\bar{R}) < I(\Omega, \sup) + \varepsilon = I(\Omega) - \varepsilon$, a contradiction establishing the result.

Lemmas 2.1, 2.2 and 2.3 together prove Theorem 1.

3. Proof of Theorem 2. Let $D_h = X_1^h \cup \cdots \cup X_h^h$ have the special form $(x_0 = -\infty, x_1] \cup (x_1, x_2] \cup \cdots \cup (x_{h-1}, x_h = \infty)$, where $P(X_i^h) = 1/h$, $i = 1, \cdots, h$. Given $\varepsilon > 0$, there exists a distribution function G_h and associated probability distribution in Ω, which we also denote G_h, with $I(G_h, D_h) - I(\Omega, D_h) < \varepsilon$. Let $\bar{G}_h$ be defined from G_h just as $\bar{R}$ was from R. Then we find:

$$s(G_h, \bar{G}_h) \leqq \max_i G_h(X_i^h) \tag{3.1}$$

$$I(\bar{G}_h) = \textstyle\sum_i G_h(X_i^h) \log [hG_h(X_i^h)] \tag{3.2}$$

$$\infty > I(\Omega) \geqq I(\Omega, D_h) > I(G_h, D_h) - \varepsilon = I(\bar{G}_h) - \varepsilon . \tag{3.3}$$

From (3.2) and (3.3), $\max_i G_h(X_i^h) \to 0$ as $h \to \infty$, whence, by (3.1),

$$s(G_h, \bar{G}_h) \to 0 \quad \text{as} \quad h \to \infty . \tag{3.4}$$

If $\bar{G}_h \in \Omega$, $I(\bar{G}_h) \geqq I(\Omega)$ and then

$$I(\Omega, D_h) > I(\Omega) - \varepsilon \tag{3.5}$$

by (3.3). If $\bar{G}_h \notin \Omega$ then $T(\bar{G}_h) < 0$ while $T(G_h) \geqq 0$, whence $|T(G_h) - T(\bar{G}_h)| \geqq |T(\bar{G}_h)|$. So, by the uniform continuity of T, (3.4) implies $|T(\bar{G}_h)| \to 0$ as $h \to \infty$. The continuity of I_r at $r = 0$, the inequality $I(\bar{G}_h) \geqq I_{T(\bar{G}_h)}$ and (3.3) then imply that (3.5) holds with ε replaced by 2ε and for h large enough. But ε is arbitrary so, with (P1), we have

$$\lim_{h\to\infty} I(\Omega, D_h) = I(\Omega)\,. \tag{3.6}$$

(Note, for comparison with Lemma 2.3, that (3.6) implies $I(\Omega, \sup) = I(\Omega)$.)

The proof of Theorem 2 is completed by the observation that, for arbitrary $\rho > 0$, the set $\mathscr{F}_\rho = \{F \mid 0 < T(F) < \rho\}$ (which is in Ω and $\{F \mid I(\Omega) \leqq I(F) < I_\rho\}$) is non-empty and open with respect to s and is included, for small ρ and large h, in the set $\{F \mid I(\Omega, D_h) \leqq I(F) < I(\Omega, D_h) + \varepsilon\}$. We may therefore choose $R \in \mathscr{F}_\rho \subset \Omega$ satisfying (C2). That (C3) holds for small enough δ follows from the inequality

$$s(Q, R) \leqq (h + 1) \max_i |Q(X_i^h) - R(X_i^h)| + \max_i R(X_i^h)\,,$$

the fact that $\sup_i R(X_i^h) \to 0$ as $h \to \infty$ (just as for G_h above) and the openness of $\mathscr{F}_\rho$.

4. Discussion. That condition (C) has limitations is illustrated by the case: $X \sim R^1$, $P \sim F_0$ (non-degenerate) with $\int x\,dF_0(x) = 0$ and $\Omega = \{F \mid \int x\,dF \geqq \alpha > 0\}$. For this case it can be shown that (C) holds if and only if either x is bounded above or $I(\Omega) = 0$. The interesting case of x unbounded and $I(\Omega) > 0$ is, however, easily dealt with by a truncation method with Theorem 1 and the result $P(F_n \in \Omega) = \exp[-nI(\Omega) + o(n)]$ is derivable without explicit dependence on the "Cramér condition" (see Bahadur and Rao (1960)).

Theorem 2 is inapplicable when x is unbounded because the required $T(F)$ is $\int x\,dF(x)$ which is not uniformly continuous. Hoadley (1967) introduced his elaborated condition to deal with such a breakdown of uniform continuity.

It is easily seen that the following otherwise general example cannot satisfy condition (C):

$$P \text{ has non-finite support and } \quad \Omega = \{Q \mid I(Q) \geqq \alpha > 0\}\,. \tag{4.1}$$

(For every D_h, we see that $I(\Omega, D_h) = 0$, which, with the choices $\varepsilon < \alpha$ and $Q = \bar{R}$, contradicts (C3).) Sanov [4] uses an indirect method to show that the distribution function version of (4.1) cannot be F-distinguishable.

The condition of Theorem 2 is the specialisation to the one-sample case of the condition of Hoadley's Theorem 1, which establishes essentially the same as our Theorem 1.

It would be of interest to know the relationship of Sanov's F-distinguishability condition, the present condition (C) and the one-sample specialisation of the elaborate version of Hoadley's condition developed for his Theorem 2.

REFERENCES

[1] BAHADUR, R. R. and RAO, R. R. (1960). On deviations of the sample mean. *Ann. Math. Statist.* **31** 1015–1027.

[2] HOADLEY, A. B. (1967). On the probability of large deviations of functions of several empirical cdf's. *Ann. Math. Statist.* **38** 360–381.

[3] HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.* **36** 369–401.

[4] SANOV, I. N. (1957). On the probability of large deviations of random variables (in Russian). *Mat. Sbornik N.S.* **42** (**84**) 11–44. (English translation in *Selected Transl. Math. Statist. Prob.* **1** (1961) 213–244.)

UNIVERSITY COLLEGE
GOWER STREET
LONDON WC1E 6BT
ENGLAND