CONVERGENCE RATES OF LARGE DEVIATIONS PROBABILITIES FOR POINT ESTIMATORS

by

Herman Rubin*

and

Andrew L. Rukhin**

Mimeograph Series #81-47

Purdue University

Department of Statistics

Division of Mathematical Sciences

September 1981

* Research supported by U.S. Army Research Office, contract number DA80K0043.

** Research supported by National Science Foundation grant MCS-8101670

Convergence Rates of Large Deviations Probabilities for Point Estimators

ABSTRACT

The convergence rates of large deviations probabilities are determined for a class of estimators of a real parameter. We also give a simple upper bound for probabilities of large deviations when the latter are measured in terms of the Chernoff function.

Key words and phrases: Probabilites of large deviations, minimum contrast estimators, M-estimators, Chernoff function.

AMS 1980 subject classifications. Primary 62F12, Secondary 62F35, 62E20,62B10.

1. Introduction and Summary

Let $\{P_\theta, \theta \in \Theta\}$ be a family of probability distributions over a space X and assume that the parameter space Θ is an open subset of the real line. If $\underline{x} = (x_1,\ldots,x_n)$ is a random sample from this family and $\delta = \delta(\underline{x})$ is a consistent estimator of θ, then the rate at which $P_\theta\{|\delta-\theta|>\varepsilon\}$ tends to zero for a fixed $\varepsilon > 0$ is of importance in large sample estimation theory. In the case of a more general parameter space with a metric D these probabilities of large deviations have the form $P_\theta\{D(\delta,\theta)>\varepsilon\}$. Useful lower bounds for the limits of these probabilities have been obtained by Bahadur (see Bahadur (1967) and Bahadur, Gupta and Zabell (1980)). Further results concerning the asymptotical behavior of large deviations probabilities for various estimators have been given by Fu (1973), (1975).

In this paper we determine the exact rate of convergence of large deviations probabilities for a class of approximate M-estimators, consistency of which has been established by Huber (1967). In Section 2 a simple upper bound is obtained for probabilities of large deviations when these are measured by means of the Chernoff function and the parameter takes values in an abstract space.

2. An Upper Bound for Large Deviations Probabilities.

Let $x_1,x_2,\ldots$ be a sequence of i.i.d. random variables wtih common distribution $P = P_{\theta_0}$ where θ_0 is a fixed parametric value. In this section we consider a class of statistics $\delta=\delta_n(\underline{x})$ such that there exists a sequence q_n, $q_n \to 0$,

$$n^{-1} \sum_1^n W(x_j,\delta) - \inf_\theta \; n^{-1} \sum_1^n W(x_j,\theta) \le q_n.$$

Here W is a measurable real function on X x Θ, and the parameter space Θ is assumed to be a locally compact Hausdorff topological space.

This class of approximate minimum contrast estimators (which of course includes maximum likelihood estimators) was considered by Huber (1967). Using the idea of original Wald's proof (1949) Huber under mild regularity conditions proved the consistency of statistics from this class (see also Pfanzagl (1969), Perlman (1972)).

Under somewhat different assumptions we show that the large deviation probabilities for these procedures tend to zero exponentially fast and give a simple upper bound for these probabilities.

Denote for any measurable set C

$$q(x,C) = \inf_{t \in C} W(x,t)$$

and

$$\rho(C) = \inf_{s>0} E \exp\{s[W(x,\theta_0) - q(x,C)]\}.$$

Our assumptions have the form:

Assumption 1. For all $\theta \neq \theta_0$

$$E[W(x,\theta_0) - W(x,\theta)] < 0.$$

Assumption 2. For any a, $0 < a < 1$ there exists a compact set A and sets D_k, $k = 1,\ldots,K$ such that $\bigcup_k D_k \supset A^c$ and

$$\rho(D_k) \leq a. \qquad k = 1,\ldots,K. \tag{2.1}$$

Assumption 3. For any θ and positive ε there exists a neighborhood B of θ such that

$$\rho(B) \leq \rho(\theta) + \varepsilon, \tag{2.2}$$

where $\rho(\theta) = \rho(\{\theta\})$ is the Chernoff function,

$$\rho(\theta) = \inf_{s>0} E \exp\{s[W(x,\theta_0) - W(x,\theta)]\}.$$

Notice that Assumption 3 is quite similar to Assumption 2 of Bahadur (1965) where it was used to prove asymptotic optimality of the likelihood ratio test.

Theorem 1, Under Assumptions 1-3 for any a, $0 < a < 1$

$$\limsup_{n\to\infty} P^{1/n}\{\rho(\delta) \le a\} \le a.$$

Proof. Since $\rho(\theta_0) = 1 > a$, the point θ_0 does not belong to the set $E = \{\theta: \rho(\theta) \le a\}$. Also the event $\delta \in E$ implies that

$$\inf_{\theta\in E} n^{-1} \sum_1^n W(x_j,\theta) \le n^{-1} \sum_1^n W(x_j,\delta).$$

Since for any fixed positive ε_0 and all sufficiently large n

$$n^{-1} \sum_1^n W(x_j,\delta) - n^{-1} \sum_1^n W(x_j,\theta_0)$$

$$\le n^{-1} \sum_1^n W(x_j,\delta) - \inf n^{-1} \sum_1^n W(x_j,\theta) < \varepsilon_0,$$

one concludes that the event $\delta \in E$ implies

$$\inf_{\theta\in E} n^{-1} \sum_1^n W(x_j,\theta) \le n^{-1} \sum_1^n W(x_j,\theta_0) + \varepsilon_0.$$

Thus for large n

$$P\{\rho(\delta) \le a\} \le P\{\inf_{\theta\in E} n^{-1} \sum_1^n W(x_j,\theta) \le n^{-1} \sum_1^n W(x_j,\theta_0) + \varepsilon_0\}. \quad (2.3)$$

Let A and D_k, $k = 1,\ldots,K$ be the sets such that (2.1) holds, and for a positive ε let B_θ be a neighborhood of θ for which (2.2) is satisfied. Then the compact set $C = A \cap E$ is covered by open sets $B_\theta, \theta\in C$. Therefore there exists a finite subcovering, say, $B_1,\ldots,B_m$ of C.

Denote for $i=1,\ldots,m$, $p = m+1,\ldots,m+K$

$$q_i(x) = q(x,B_i),$$

$$q_p(x) = q(x,D_{p-m}).$$

One has with $M = m + K$

$$\inf_{\theta\in E} n^{-1} \sum_1^n W(x_j,\theta) \ge \min_{1\le i\le M} n^{-1} \sum_1^n q_i(x_j).$$

It follows from (2.3)

$$P\{\delta\in E\} \le P\{\min_{1\le i\le M} n^{-1} \sum_{1}^{n} [W(x_j,\theta_0) - q_i(x_j)] + \varepsilon_0 > 0\}$$

$$\le \sum_{i=1}^{M} P\{n^{-1} \sum_{1}^{n} [W(x_j,\theta_0) - q_i(x_j)] + \varepsilon_0 > 0\}$$

$$\le M \max_{1\le i\le M} P\{n^{-1} \sum_{1}^{n} [W(x_j,\theta_0) - q_i(x_j)] + \varepsilon_0 > 0\}.$$

Because of Chernoff's Theorem (see Chernoff (1952))

$$P^{1/n}\{n^{-1} \sum_{1}^{n}[W(x_j,\theta_0) - q_i(x_j)] + \varepsilon_0 > 0\}$$

$$\to \inf_{s>0} e^{s\varepsilon_0} E \exp\{s[W(X,\theta_0) - q_i(X)]\},$$

so that

$$\limsup P^{1/n}\{\delta\in E\} \le \max_{1\le i\le M} \inf_{s>0} e^{s\varepsilon_0} E \exp\{s[W(X,\theta_0) - q_i(X)]\}. \quad (2.4)$$

Notice that for a fixed i, $i = 1,\ldots,M$ and any positive ε_1 there exists s_1 such that

$$E \exp\{s_1[W(X,\theta_0) - q_i(X)]\} < \inf_{s>0} E \exp\{s[W(X,\theta_0) - q_i(X)]\} + \varepsilon_1.$$

Therefore for sufficiently small ε_0

$$\inf_{s>0} e^{s\varepsilon_0} E \exp\{s[W(X,\theta_0) - q_i(X)]\} \le e^{s_1\varepsilon_0} E \exp\{s_1[W(X,\theta_0) - q_i(X)]\}$$

$$\le \inf_{s>0} E \exp\{s[W(X,\theta_0) - q_i(X)]\} + 2\varepsilon_1.$$

Thus in (2.4) we can let ε_0 tend to zero and obtain

$$\limsup P^{1/n}\{\delta \in E\} \le \max_{1\le i\le M} \inf_{s>0} E \exp\{s[W(X,\theta_0) - q_i(X)]\}.$$

Assumption 3 entails for i=1,...,m

$$\inf_{s>0} E \exp\{s[W(X,\theta_0) - q_i(X)]\} \le \rho(\theta_i) + \varepsilon \le a + \varepsilon,$$

and because of Assumption 2 for $i = m + 1,\dots,M$

$$\inf_{s>0} E \exp\{s[W(X,\theta_0) - q_i(X)]\} \le a.$$

Since ε was an arbitrary positive number Theorem 1 is proven.

Remark 1. For smooth functions W the probability $P^{1/n}(\rho(\delta) \le a)$ for small a typically behaves as a^c with $c \ge 2$. However examples with discrete parameter θ or nondifferentiable functions W show that the bound of Theorem 1 is the best possible without any regularity assumptions (also see Section 3).

Remark 2. Assumptions 2 and 3 can be modified and weakened. For instance the function $\rho(C)$ can be replaced by

$$\left[\inf_{s>0} E \exp\{s \sum_1^m [W(x_j,\theta_0) - q\,(x_j,C)]\}\right]^{1/m},$$

where m is an fixed positive integer and $(x_1,\dots,x_m)$ are i.i.d. random variables with distribution P. (See Perlman (1972)). These modifications (with $m\ge 2$) allow to establish Theorem 1 for location-scale parameter families under mild moment conditions.

3. Convergence Rates for Approximate M-estimators.

In this section we determine the exact convergence rate for some approximate M-estimators of a real parameter θ. These estimators δ satisfy the following condition: there exists a (nonrandom) sequence q_n, $q_n \to 0$ such that

$$|n^{-1} \sum_1^n w(x_j,\delta)| \le q_n.$$

Here $w(x,\theta)$ is a real function over $X \times \mathbb{R}$. The consistency of approximate M-estimators has been also established by Huber (1967).

<u>Theorem 2</u>. <u>Assume that for each fixed</u> x, $w(x,\cdot)$ <u>is a decreasing function, and let</u> $\delta = \delta_n(\underline{x})$ <u>be the corresponding approximate M-estimator. If for a positive</u> ε

$$P_\theta\{w(X,\theta+\varepsilon) > 0\} > 0, \tag{3.1}$$

<u>then</u>

$$P_\theta^{1/n}\{\delta > \theta + \varepsilon\} \to \inf_{s>0} E_\theta \exp\{sw(X,\theta+\varepsilon)\} = e_1(\theta,\varepsilon), \tag{3.2}$$

$$P_\theta^{1/n}\{\delta < \theta - \varepsilon\} \to \inf_{s>0} E_\theta \exp\{-sw(X,\theta-\varepsilon)\} = e_2(\theta,\varepsilon) \tag{3.3}$$

<u>and</u>

$$P_\theta^{1/n}\{|\delta-\theta| > \varepsilon\} \to \max[e_1(\theta,\varepsilon),\ e_2(\theta,\varepsilon)]. \tag{3.4}$$

<u>Proof.</u> One has for arbitrary small positive ε_0

$$P_\theta\{\delta > \theta+\varepsilon\} \le P_\theta\{ n^{-1} \sum_1^n w(x_j,\eta) > -\varepsilon_0 \text{ for } \eta \le \theta + \varepsilon\}$$

$$= P_\theta\{n^{-1} \sum_1^n w(x_j,\theta + \varepsilon) > -\varepsilon_0\}.$$

Also

$$P_\theta\{n^{-1} \sum_1^n w(x_j,\theta+\varepsilon) > \varepsilon_0\} \le P_\theta\{\delta > \theta+\varepsilon\}.$$

Applying Chernoff's Theorem we obtain

$$\inf_{s>0} e^{-s\varepsilon_0} E_\theta \exp\{sw(X,\theta+\varepsilon)\} \le \lim P_\theta^{1/n}\{\delta > \theta+\varepsilon\}$$

$$\le \inf_{s>0} e^{s\varepsilon_0} E_\theta \exp\{sw(X,\theta+\varepsilon)\} \tag{3.5}$$

Because of (3.1) the functions of ε_0 in (3.5) are continuous. Since ε_0 was arbitrary one deduces

$$\lim P_\theta^{1/n} \{\delta > \theta + \varepsilon\} = \inf_{s>0} E_\theta \exp\{sw(X,\theta+\varepsilon)\} .$$

Analogously formula (3.3) is established which together with (3.2) implies (3.4).

Remark 1. It follows from Theorem 2 that the probabilities of large deviations $P_\theta\{|\delta-\theta| > \varepsilon\}$ tend to zero exponentially fast if and only if both infima in (3.4) are smaller than one, i.e.

$$E_\theta\, w(X,\theta + \varepsilon) < 0$$

and

$$E_\theta\, w(X,\theta - \varepsilon) > 0.$$

These conditions are met of course when

$$E_\theta\, w(X,\theta) = 0 \tag{3.6}$$

and $w(x,\theta)$ is strictly monotone. In the case when $w(x,\theta) = \log' p(x,\theta)$, wehre $p(x,\theta)$ is the density of P_θ, (so that approximate maximum likelihood estimator obtains) condition (3.6) is satisfied under mild regularity assumptions. Monotonicity of $w(x,\theta)$ means the log-concavity of $p(x,\theta)$.

Remark 2. Notice that Theorem 2 does not hold if the function w is not monotone. This can be seen by the example where w is the logarithmic derivative of Cauchy distribution density with location parameter θ. In this case probabilities of large deviations for the maximum likelihood estimator do not have the rate given by the right-hand side of (3.4). However if the latter is a decreasing function of ε and assumptions of Theorem 1 are met then (3.4) provides an upper bound for probabilities of large deviations.

In the case of Cauchy distribution this monotonicity does not hold but Theorem 1 shows that probability of large deviations decrease exponentially. Their precise rate is unknown and its determination seems to be a difficult problem.

Remark 3. Note that one can derive from Theorem 2 a class of estimators $\hat{\delta}$ which are asymptotically efficient after Bahadur:

$$\lim_{\varepsilon\to 0} \lim_{n\to\infty} n^{-1} \varepsilon^{-2} \log P_\theta\{|\hat{\delta} - \theta| > \varepsilon\} = -I(\theta)/2. \tag{3.7}$$

Indeed (3.7) holds if $\hat{\delta}$ is an approximate M-estimator with $w(x,\theta) = \log' p(x,\theta)$, which is easy to check by taking limit as ε tends to zero for the logarithm of the right-hand side of (3.4). An analogue of the result of Fu (1973) concerning asymptotical efficiency of the generalized maximum likelihood estimator $\hat{\delta}$,

$$\prod_1^n p(x_j, \hat{\delta})\lambda(\hat{\delta}) = \max_\theta \prod_1^n p(x_j,\theta)\lambda(\theta),$$

can be obtained from this remark since $\hat{\delta}$ is an approximate M-estimator with $w(x,\theta) = \log' p(x,\theta)$ (the last function is assumed to be monotone and the function $\log'\lambda$ is assumed to be bounded).

Remark 4. Assume that θ is real location parameter, and let δ be the Pitman estimator in the problem of confidence estimation of θ by means of an interval of width 2ε, i.e.

$$\prod_i^n p(x_j - \delta - \varepsilon) = \prod_i^n p(x_j - \delta + \varepsilon).$$

If $w(x,\theta) = w(x-\theta)$ and δ_1 is the corresponding equivariant M-estimator, then the comparison of δ and δ_1 shows that

$$\max[\inf_{s>0} \int p^{1-s}(x)p^s(x \pm 2\varepsilon)dx]$$

$$\leq \max[\inf_{s>0} \int \exp\{sw(x \pm \varepsilon)\} p(x)dx].$$

It is assumed here that log p is a concave function (i.e. the family $\{p(\cdot - \theta)\}$ has a monotone likelihood ratio) and that w is a monotone function.

Remark 5. Theorem 2 holds if the approximate M-estimators δ are defined as follows: for any measurable set E and a positive ε_0

$$\{n^{-1} \sum_{1}^{n} w(x_j,t) > \varepsilon_0, \quad t \in E\} \subset \{\delta \notin E\}$$

$$\subset \{n^{-1} \sum_{1}^{n} w(x_j,t) > -\varepsilon_0, \; t \in E\},$$

if n is sufficiently large.

As an application of Theorem 2 let us consider the situation when

$$w(x,\theta) = x-\psi(\theta)$$

where ψ is an increasing function. Clearly the corresponding M-estimator δ satisfies the condition

$$\psi(\delta) = \bar{x} = \sum_{1}^{n} x_j/n,$$

and

$$P_\theta(\delta > \theta + \varepsilon) = P_\theta(\bar{x} > \psi(\theta+\varepsilon)) = P_\theta(\bar{x} \geq a).$$

Thus Theorem 2 contains Chernoff's Theorem,

$$P_\theta^{1/n}(\bar{x} \geq a) \to \inf_{s>0} E_\theta \exp\{s(X-a)\},$$

as a particular case.

In another example let $w(x,\theta) = -1,\ x < \theta;\ = 1,\ x > \theta$. Then the corresponding M-estimator can be interpreted as the median, $\delta = x_{(1/2)}$, and

$$P_\theta^{1/n}(x_{(1/2)} > \theta + \varepsilon) \to \inf_{s>0}[e^s P_\theta(X > \theta + \varepsilon) + e^{-s} P_\theta(X < \theta + \varepsilon)]$$

$$= [2P_\theta(X < \theta + \varepsilon) P_\theta(X > \theta + \varepsilon)]^{1/2}.$$

For instance, P is a double exponential distribution, then according to this formula

$$P_\theta^{1/n}(x_{(1/2)} > \theta + \varepsilon) \to e^{-|\theta|/2}(2-e^{-|\theta|})^{1/2} = \sigma(\theta).$$

Since for $W(x,\theta) = |x-\theta| + \log 2$

$$\rho(\theta) = \inf_{\theta>0} E\exp\{s[W(x,0)-W(x,\theta)]\}$$

$$= e^{-|\theta|/2}(1+|\theta|/2),$$

Theorem 1 implies that

$$\lim_{n\to\infty} P_\theta^{1/n}(|x_{(1/2)}| > t) \leq \rho(t).$$

This inequality means

$$(2-e^{-|\theta|})^{1/2} \leq (1+|\theta|/2).$$

In this example, $\log \sigma(\theta)/\log \rho(\theta)$ decreases from 4 as $|\theta| \to 0$ to 1 as $|\theta| \to \infty$. This shows that the bound in Theorem 1 cannot be improved.

REFERENCES

[1] Bahadur, R.R. (1965), An optimal property of the likelihood ratio statistic. Proc. Fifth Berkeley Symp. Math. Statist. Prob., 1, 13-26.

[2] Bahadur, R.R. (1967). Rates of convergence of estimates and test statistics. Ann. Math. Statist. 38 304-324.

[3] Bahadur, R.R., Gupta, J.C. and Zabell, S.L. (1980). Large deviations, tests, and estimates, in "Asymptotic Theory of Statistical Tests and Estimation", edited by I.M. Chakravarti, Academic Press, 33-64.

[4] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann. Math. Statist. 23, 493-507.

[5] Fu, J.C. (1973). On a theorem of Bahadur on the rate of convergence of point estimators. Ann. Statist. 1, 745-749.

[6] Fu, J.. (1975). The rate of convergence of consistent point estimators, Ann. Statist. 3, 234-240.

[7] Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1, 221-233.

[8] Perlman, M.D. (1972). On strong consistency of approximate maximum likelihood estimators, Proc. Seventh Berkeley Symp. Math. Statist. Prob. 1, 263-281.

[9] Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimators. Metrika 19, 249-272.

[10] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist. 20, 595-601.