How many moments can be estimated from a large sample?

Abram Kagan\textsuperscript{a,}\textsuperscript{*}, Sergei Nagaev\textsuperscript{b}

\textsuperscript{a}Department of Mathematics, University of Maryland, College Park, Mathematics Building 1301, MD 20742, USA
\textsuperscript{b}Mathematics Institute, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia

Received November 2000; received in revised form April 2001

Abstract

Under general conditions on a population, it is proved that the number of population moments that can be simultaneously consistently estimated by the corresponding sample moments in a large sample of size \(n\) is roughly \((\log n)/(2\log\log n)\) and this order is rather sharp. © 2001 Elsevier Science B.V. All rights reserved

MSC: primary 62F12

Keywords: Large sample; Moments; Consistency

1. Introduction

Let \(X_1,\ldots,X_n\) be a sample of size \(n\) from a univariate population \(F\) with all the moments finite

\[
\alpha_k = \int x^k \, dF(x), \quad k = 1, 2, \ldots
\]

Let

\[
\mu_k = \int (x - \mu_1)^k \, dF(x), \quad k = 2, 3, \ldots
\]

denote the central moments of \(F\).

By virtue of the law of large numbers, the sample moment

\[
a_{n,k} = (1/n)(X_1^k + \cdots + X_n^k)
\]

is a consistent (actually, strongly consistent by virtue of the SLLN) estimator of the population moment \(\alpha_k\) for any fixed \(k\), i.e., for any given \(\epsilon > 0\)

\[
P(|a_{n,k} - \alpha_k| > \epsilon) \to 0, \quad n \to \infty.
\]
Moreover, the vector
\[
(\sqrt{n}(a_{n,1} - \alpha_1), \ldots, \sqrt{n}(a_{n,k} - \alpha_k))
\]
converges in distribution to a Gaussian vector with mean zero and covariance matrix
\[
V = (v_{ij}), \quad v_{ij} = \alpha_{i+j} - \alpha_i\alpha_j, \quad i, j = 1, \ldots, k
\]
(see, e.g., Cramér, 1946, Chapter 28).

One can easily see from this that the normalized vector \((\sqrt{n}(\bar{X} - \alpha_1), \sqrt{n}(S^2 - \mu_2))\) of the sample mean and sample variance converges in distribution to a bivariate Gaussian vector with mean zero and covariance matrix
\[
V = \begin{pmatrix}
\mu_2 & \mu_3 \\
\mu_3 & \mu_4 - \mu_2^2
\end{pmatrix}.
\]

In particular, if \(\mu_2 = 0\) then \(\bar{X}\) and \(S^2\) are asymptotically independent. This example is often used to illustrate differences between large and small sample properties. As is well known (see, e.g., Kagan et al., 1973, Theorem 4.2.1), in small samples (i.e., for finite \(n\)) \(\bar{X}\) and \(S^2\) are independent iff \(F\) is Gaussian while in large samples (i.e., asymptotically as \(n \to \infty\)) they are independent iff the asymmetry of \(F\) is zero.

Since from (3) it follows that for any fixed integer \(k\) and \(a > 0\)
\[
P\left\{\max_{1 \leq j \leq k} |a_{n,j} - \alpha_j| > a\right\} \leq \sum_{j=1}^{k} P\{|a_{n,j} - \alpha_j| > a\} \to 0, \quad n \to \infty, \tag{4}
\]
there exists a function \(k_n = k_n(F) \to \infty\) as \(n \to \infty\) such that (4) holds for \(k = k_n\).

In other words, one can simultaneously consistently estimate the number of population moments that goes to infinity together with the sample size.

We show in Section 2 that if \(F\) satisfies the condition
\[
|\alpha_k| \leq k!H^k, \quad k = 1, 2, \ldots \tag{5}
\]
for some constant \(H\), then (4) holds for
\[
k = k_n = \frac{(1 - \delta)\log n}{2\log\log n}(1 + o(1)), \quad n \to \infty, \tag{6}
\]
where \(\delta\) is an arbitrary positive constant.

Inequalities (5) are actually the well known Bernstein condition used in proving the exponential inequality bearing his name (Bernstein, 1946; see also Petrov, 1995, Chapter 3, Theorem 18). The Bernstein condition is equivalent to analyticity of the characteristic function \(f(t) = \int e^{itx}dF(x)\) in the strip \(-1/H < T(t) < 1/H\).

If \(F\) satisfies (5) we shall write \(F \in \mathcal{B}\).

Order (6) of the number of simultaneously estimable moments is rather sharp if the population moments are estimated by their sample counterparts. We show that if for \(F \in \mathcal{B}\)
\[
|\alpha_k| = k!H^k(1 + o(1)), \quad k \to \infty \tag{7}
\]
then for
\[
k = K_n(F) = \frac{(1 + \delta)\log n}{2\log\log n}(1 + o(1)), \quad n \to \infty \tag{8}
\]
with an arbitrary $\delta > 0$

$$P \left\{ \max_{1 \leq j \leq K_n} |a_{n,j} - x_j| > \varepsilon \right\} \to 1$$

for any $\varepsilon > 0$.

Our result is a warning, in a sense; one should be very careful in using too many sample moments even when the sample size is rather large. The function $l(n) = (\log n)/(2 \log \log n)$ is growing very slowly:

$$l(1,000) = 1.79, \quad l(10,000) = 2.07, \quad l(100,000) = 2.36, \quad l(1,000,000) = 2.63.$$  

It is also interesting to compare the result with the behavior of the empirical characteristic function $\hat{f}_n(t) = \frac{1}{n}(e^{itx_1} + \cdots + e^{itx_n})$ as an estimator of the population characteristic function $f(t)$. For any fixed $t$, $\hat{f}_n(t)$ is a strongly consistent estimator of $f(t)$. How long is the interval $[-T_n, T_n]$ where

$$\max_{-T_n < t < T_n} |\hat{f}_n(t) - f(t)| \to 0$$

in probability (or with probability 1)?

Csörgő and Totik (1983) proved (even for multivariate distributions) that if

$$\lim_{n \to \infty} \log T_n = 0$$

(equivalently, $T_n = \exp\{o(n)\}$) then (9) holds with probability 1.

If

$$\lim_{n \to \infty} \sup_{n} \frac{\log T_n}{n} > 0$$

then (9) does not hold even in probability for any $F$ such that $f(t) \to 0$ as $|t| \to \infty$.

Thus, the empirical characteristic function is uniformly consistent on “almost exponentially” growing intervals.

It is instructive to compare the phenomenon discussed in this paper, namely, the lack of uniform consistency of $(1 + \delta)(\log n)/(2 \log \log n)$ sample moments for sampling from distributions $F$ satisfying (7), with what sometimes occurs in nonparametric statistics.

Let $\psi(F)$ be a functional of interest defined on a set $\mathcal{F}$ of distributions $F$. Suppose that $\hat{\psi}_n(x_1, \ldots, x_n)$ is a consistent estimator of $\psi(F)$ constructed from a sample from $F$, i.e.,

$$P_F \{ |\hat{\psi}_n - \psi(F)| > \varepsilon \} \to 0, \quad n \to \infty$$

for any $F \in \mathcal{F}$. (The subindex $F$ in $P$ emphasizes the fact that the sample is drawn from $F$.) However, sometimes (e.g., in estimating probability density function in $L^1$-norm for $\mathcal{F}$ big enough; see Devroye and Györfi (1985, Chapter 4))

$$\sup_{F \in \mathcal{F}} \{ |\hat{\psi}_n - \psi| > \varepsilon \} \geq A$$

for all $n$.

Let now $\psi^{(1)}(F), \psi^{(2)}(F), \ldots$ be an infinite series of functionals of interest estimated by $\hat{\psi}_n^{(1)}, \hat{\psi}_n^{(2)}, \ldots$, respectively. Though each $\hat{\psi}_n^{(k)}$ is a consistent estimator of $\psi^{(k)}(F)$, it may well occur that if $k = k_n \to \infty$ fast
enough, then
\[ P_F \left\{ \max_{1 \leq j \leq k} |\psi_{n,j}^{(j)} - \psi^{(j)}(F)| > \varepsilon \right\} \geq \Delta \] (11)
for all \( \varepsilon > 0 \) and an individual \( F \).

Thus, (10) occurs due to the fact that \( \mathcal{F} \) is big while (11) occurs due to the number of estimated functionals.

In this paper, a special case of (11) is discussed when \( \psi^{(j)}(F) = \int x^j \text{d}F \).

2. Result

**Theorem 1.** Let \((X_1, \ldots, X_n)\) be a sample from distribution \( F \in \mathcal{B} \). For \( k_n = (1 - \delta) \log n / 2 \log \log n (1 + o(1)) \) for some \( \delta > 0 \), then for any \( \varepsilon > 0 \),
\[ \lim_{n \to \infty} P \left\{ \max_{1 \leq j \leq k_n} |a_{n,j} - z_j| < \varepsilon \right\} = 1. \] (12)

If \( K_n = (1 + \delta) \log n / (2 \log \log n) (1 + o(1)) \) for some \( \delta > 0 \) and \( F \in \mathcal{B} \) satisfies (7), then
\[ \lim_{n \to \infty} P \left\{ \max_{1 \leq j \leq k_n} |a_{n,j} - z_j| > C \right\} = 1 \] (13)
for any \( C > 0 \).

In other words, for samples from any \( F \in \mathcal{B} \) the first \( k_n \) sample moments \( a_j, j = 1, \ldots, k_n \) converge in probability to the corresponding population moments uniformly in \( j \). And for samples from any \( F \in \mathcal{B} \) satisfying (7) (there are many such \( F \); one example is \( F(x) = 1 - e^{-x}, x > 0 \) for which \( z_k = k! \), \( k = 1, 2, \ldots \)), the first \( K_n \) sample moments do not converge to their population counterparts uniformly.

Condition (7) is crucial for the second part of Theorem 1. Without it, the number of moments that are simultaneously consistently estimable may be of higher order than \( (\log n) / (2 \log \log n) \) and even infinite. Take, for example, an arbitrary distribution \( F \) concentrated on \((0, b), 0 < b < 1\). For \( X \sim F \), \( \text{var}(X^j) \leq E(X^{2j}) = b^{2j+1} / (2j+1) \), \( j = 1, 2, \ldots \). For a sample from distribution \( F \), one has by virtue of the Chebyshev’s inequality,
\[ P \left\{ \max_{1 \leq j < \infty} |a_{n,j} - z_j| > \varepsilon \right\} \leq \sum_{j=1}^{\infty} P\{|a_{n,j} - z_j| > \varepsilon\} \leq \frac{1}{n\varepsilon^2} \sum_{j=1}^{\infty} b^{2j+1} / (2j+1) \leq \text{const} / (n\varepsilon^2) \rightarrow 0, \quad n \to \infty \]
for any \( \varepsilon > 0 \).

For a sample from uniform \((0, 1)\) distribution \( F \), a similar rough estimate shows that if \( k_n = \exp\{o(n)\} \), then \( k_n \) first sample moments uniformly converge in probability to the corresponding moments of \( F \).

**Proof of Theorem 1.** For any \( \varepsilon > 0 \) one has by virtue of Chebyshev’s inequality
\[ P \left\{ \max_{1 \leq j \leq k} |a_{n,j} - z_j| > \varepsilon \right\} \leq \sum_{j=1}^{k} P\{|a_{n,j} - z_j| > \varepsilon\} \leq (1 / \varepsilon^2) \sum_{j=1}^{k} \text{var}(a_{n,j}) = (1 / n\varepsilon^2) \sum_{j=1}^{k} (z_{2j} - z_j^2) \leq (1 / n\varepsilon^2) \sum_{j=1}^{k} z_{2j}. \] (14)
For $\bar{H} = \max(1, H)$ it follows from (5)

$$a_{2j} \leq (2j)!\bar{H}^{2j}$$

whence $a_{2j} \leq (2k)!\bar{H}^{2k}$, $j = 1, \ldots, k$ and

$$\sum_{j=1}^{k} a_{2j} \leq k(2k)!\bar{H}^{2k}.$$  

(15)

By virtue of the Stirling formula

$$N! = \sqrt{2\pi N} (N/e)^N (1 + o(1)), \quad N \to \infty,$$

one gets from (14) and (15)

$$(1/n^2) \sum_{j=1}^{k} a_{2j} \leq (1/n^2) \exp\{2k \log k - \log n + o(k \log k)\}, \quad k \to \infty.$$  

(16)

If

$$k = \frac{(1 - \delta) \log n}{2 \log \log n} (1 + o(1))$$

with $\delta > 0$, then

$$\log k = \log \log n - \log \log \log n + \text{const} + o(1)$$

and

$$2k \log k - \log n + o(k \log k) = - \delta \log n + o(\log n) \to -\infty$$

so that

$$P\left\{ \max_{1 \leq j \leq k_n} |a_{n,j} - a_j| > \varepsilon \right\} \to 0, \quad n \to \infty,$$  

proving the first part of the theorem.

Turn now to the second part of Theorem 1. Since for any $C > 0$

$$P\left\{ \max_{1 \leq j \leq k_n} |a_{n,j} - a_j| > C \right\} \geq P\{ |a_{n,k} - a_k| > C \}$$

(18)

suffice to show that for the largest even $k$ not exceeding $K_n$, say $k = K_n$, the right hand side of (18) goes to 1 as $n \to \infty$.

For even $k$, $E(|X^k|) = E(X^k)$ and for $F$ satisfying (7),

$$E(X^k) = a_k = k!H^{2k}(1 + o(1)), \quad k \to \infty$$

and

$$\sigma_k^2 = \text{var}(X^k) = \{(2k)! - (kt)^2\}H^{2k}(1 + o(1)) = (2k)!H^{2k}(1 + o(1)), \quad k \to \infty$$

and

$$\beta_k = E(|X^k - a_k|^3) \leq E((X^k - a_k)^3) = (3k)!H^{3k}(1 + o(1)), \quad k \to \infty.$$
For
\[ k = K_n = \frac{(1 + \delta) \log n}{2 \log \log n} (1 + o(1)), \quad \delta > 0, \]  

(19)

\[ \log K_n = o(\log n), \quad n \to \infty \text{ and, since } K_n - K_n' \leq 2, \quad \log K_n' = o(\log n). \]

To estimate the right hand side of (18) we shall use the central limit theorem for arrays. Let us check that with \( k = K_n' \) the array of \( X_1^k, \ldots, X_n^k \), \( n = 1, 2, \ldots \), satisfies the Lyapunov condition. Since \( X_1, \ldots, X_n \) are identically distributed the Lyapunov ratio is

\[ L_n = \frac{n!}{(n\sigma_i^2)^{3/2}} = \frac{n(3k)!}{n^{3/2}(2k)!^{3/2}} (1 + o(1)). \]

The Stirling formula gives

\[ \log L_n = \text{const} - (1/2) \log k - (1/2) \log n + 3k \log(3/2) + o(1) \]

whence

\[ \log L_n = - (1/2) \log n + o(\log n) \]

and

\[ L_n \to 0, \quad n \to \infty. \]

Thus, the central limit theorem holds (see, e.g., Chung, 1974 or Shao, 1999, Theorem 1.15) and

\[ P \left\{ \frac{1}{\sigma_k \sqrt{n}} \sum_{i=1}^{n} (X_i^k - x_k) < x \right\} \to \Phi(x), \quad n \to \infty, \]

(20)

uniformly in \( x, \Phi(x) \) being the standard normal distribution function.

For any \( C > 0 \),

\[ P\{|a_n,k - x_k| > C\} = P\left\{ \frac{1}{n} \sum_{i=1}^{n} (X_i^k - x_k) > C \right\} = P\left\{ \left| \frac{1}{\sigma_k \sqrt{n}} \sum_{i=1}^{n} (X_i^k - x_k) \right| > \frac{C}{\sigma_k} \sqrt{n} \right\}. \]

By using again the Stirling formula, one gets for \( k = K_n' \)

\[ \log(\sqrt{n}/\sigma_k) = (\frac{1}{2}) \left\{ \log n - \log(2k)! - k \log H \right\} + o(1) \]

\[ = (\frac{1}{2}) \left\{ \log n - 2k \log k + 2k - \left( \frac{1}{2} \right) k \log k - k \log H \right\} + O(1) \]

\[ = (\frac{1}{2}) \log n - \left( (1 + \delta)/2 \right) \frac{\log n}{\log \log n} (\log \log n - \log \log n) + o(\log n) \]

\[ = -(\delta/2) \log n + o(\log n) \to -\infty, \quad n \to \infty, \]

(21)

so that \( C\sqrt{n}/\sigma_k \to 0 \) as \( n \to \infty. \)

Hence for any given \( \varepsilon > 0 \)

\[ \lim_{n \to \infty} P\{|a_n,k - x_k| > C\} \geq \lim_{n \to \infty} P\left\{ \frac{1}{\sigma_k \sqrt{n}} \sum_{i=1}^{n} (X_i^k - x_k) > \varepsilon \right\} = 2\Phi(-\varepsilon), \]

proving the second part of Theorem 1. \( \square \)
Acknowledgements

The work was done when the second author visited the Mathematics Department of the University of Maryland, College Park. He is grateful to the Department and especially to Mark Freidlin for arranging the visit. The authors are grateful to an anonymous referee whose comments helped improving the presentation. In particular, the last part of Introduction owes the referee its appearance.

References