AN EDGEWORTH EXPANSION FOR FINITE POPULATION U-STATISTICS

M. Bloznelis[1] F. Götze[1]

Vilnius University and IMI
University of Bielefeld

January 1997

Abstract. Suppose that U is a U-statistic of degree two based on N random observations drawn without replacement from a finite population. For the distribution of a standardized version of U we construct an Edgeworth expansion with remainder $O(N^{-1})$ provided that the linear part of the statistic satisfies a Cramér type condition.

1. Introduction and results

Let $\mathcal{A}$ denote a population of n elements $a_1, \dots, a_n$ and let $\mathcal{H} : \mathcal{A} \times \mathcal{A} \to R$ denote symmetric function of its two arguments. By $X_1, \dots, X_N$, $N \le n$ we denote random variables with values in $\mathcal{A}$ such that $X = \{X_1, \dots, X_N\}$ represents a random sample from $\mathcal{A}$ of size N drawn without replacement, i.e., $\mathbf{P}\{X \in B\} = \binom{n}{N}^{-1}$ for any subset B of $\mathcal{A}$ of size N. Assume that $\mathbf{E}\,\mathcal{H}(X_1, X_2) = 0$.

We shall investigate the asymptotic distribution of the statistic

$$U = \sum_{1 \le i < j \le N} \mathcal{H}(X_i, X_j).$$

We shall decompose the statistic as

$$U = L + Q, \tag{1.1}$$

[1]Research supported by the SFB 343 in Bielefeld.

1991 *Mathematics Subject Classification.* Primary 62E20; secondary 60F05.

Key words and phrases. Edgeworth expansion, finite population, U-statistic, Central Limit Theorem, sampling without replacement.

where

$$L=\sum_{i=1}^{N} g_1(X_i) \qquad \text{respectively} \qquad Q=\sum_{1\le i<j\le N} g_2(X_i, X_j)$$

is the linear respectively the quadratic part of the statistic. Here

$$g_1(x)=(N-1)\,t(x), \qquad t(x)=\frac{n-1}{n-2}\,\mathbf{E}\big(\mathcal{H}(X_1,\,X_2)\,|\,X_1=x\big)$$

and

$$g_2(x_1,\;x_2)=\mathcal{H}(x_1,\,x_2)-t(x_1)-t(x_2).$$

Since

$$\mathbf{E}\big(g_2(X_1,\,X_2)\,|\,X_1=x\big)=0, \qquad \forall x\in\mathcal{A}, \tag{1.2}$$

the random variables $g_1(X_i)$ and $g_2(X_j,\,X_k)$, $1\le i\le N$, $1\le j<k\le N$, (and thus the parts L and Q) are uncorrelated.

If the linear part dominates the statistic for large N, the distribution of U can be approximated by a Gaussian distribution using the Central Limit Theorem (CLT).

The asymptotic normality of linear statistics based on samples drawn without replacement from finite populations has been studied by a number of authors. Erdös and Rényi (1959) proved the CLT under very mild conditions. The rate of convergence in the CLT was first studied by Bikelis (1969). Berry–Esseen bounds of order $O(N^{-1/2})$ were obtained by Höglund (1978). Robinson (1978) proved the validity of an Edgeworth expansion with a remainder of order $O(N^{-3/2})$, see also Bickel and van Zwet (1978).

Nandi and Sen (1963) studied the behavior of finite population U-statistics and showed that under proper regularity conditions the sequence of distributions of normalized U-statistics converges to the standard normal distribution. The rate of this convergence was investigated by Zhao and Chen (1987, 1990), Kokic and Weber (1990, 1991) and, as a particular case of the rate of convergence of general multivariate sampling statistics, by Bolthausen and Götze (1993). In contrast to the case of independent and identically distributed observations where the second order asymptotic theory has been developed for U-statistics, see Bickel (1974), Götze (1979), Callaert, Janssen and Veraverbeke (1980), Bickel, Götze and van Zwet (1986), and for more general asymptotically normal symmetric statistics, see Bentkus, Götze and van Zwet (1994), there are only a few results concerned with higher order asymptotics of nonlinear finite population statistics. Babu and Singh (1985) proved the validity of an Edgeworth expansion with a remainder $o(N^{-1/2})$ for second order polynomials of variables $X_1,\dots,X_N$. Kokic and Weber (1990) established one term Edgeworth expansion with the remainder $o(N^{-1/2})$ for finite population U-statistics of degree 2.