AN EDGEWORTH EXPANSION FOR SYMMETRIC FINITE POPULATION STATISTICS

M. Bloznelis[1], F. Götze

Vilnius University and Bielefeld University

2000

Abstract. Let T be a symmetric statistic based on sample of size N drawn without replacement from a finite population of size n, where $n > N$. Assuming that the linear part of Hoeffding's decomposition of T is nondegenerate we construct one term Edgeworth expansion for the distribution function of T and prove the validity of the expansion with the remainder $O(1/N^*)$ as $N^* \to \infty$, where $N^* = \min\{N, n - N\}$.

1. Introduction and results

1. Introduction. Given a set $\mathcal{X} = \{x_1, \dots, x_n\}$, let $(X_1, \dots, X_n)$ be a random permutation of the ordered set $(x_1, \dots, x_n)$, which is uniformly distributed over the class of permutations. Let

$$T = t(X_1, \dots, X_N)$$

denote a symmetric statistic of the first N observations $X_1, \dots, X_N$, where $N < n$. That is, t is a real function defined on the class of all subsets $\{x_{i_1}, \dots, x_{i_N}\} \subset \mathcal{X}$ of size N and we assume that $t(x_{i_1}, \dots, x_{i_N})$ is invariant under permutations of its arguments. Since $X_1, \dots, X_N$ represents a sample drawn without replacement from the population $\mathcal{X}$, we call T a finite population symmetric statistic.

We shall consider statistics which are asymptotically normal when N^* and n tend to ∞. In the simplest case of linear statistics the asymptotic normality was established by Erdős and Rényi (1959) under fairly general conditions. The rate

[1]Research supported by the A.v. Humboldt Foundation

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

Key words and phrases. Edgeworth expansion, finite population, Hoeffding decomposition, sampling without replacement.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

in the Erdős-Rényi central limit theorem was studied by Bikelis (1972). Höglund (1978) proved the Berry–Esseen bound. An Edgeworth expansion was established by Robinson (1978), see also Bickel and van Zwet (1978), Schneller (1989), Babu and Bai (1996).

Asymptotic normality of nonlinear statistics was studied by Nandi and Sen (1963), who proved a central limit theorem for U statistics. The accuracy of the normal approximation of U statistics was studied by Zhao and Chen (1987, 1990), Kokic and Weber (1990). A general Berry–Esseen bound for combinatorial multivariate sampling statistics (including finite population U statistics) was established by Bolthausen and Götze (1993). Rao and Zhao (1994), Bloznelis (1999) constructed Berry-Esseen bounds for Student's t statistic. One term asymptotic expansions of nonlinear statistics, which can be approximated by smooth functions of (multivariate) sample means have been shown by Babu and Singh (1985), see also Babu and Bai (1996). For U–statistics of degree two one term Edgeworth expansions were constructed by Kokic and Weber (1990). Bloznelis and Götze (1999 a,b) established the validity of one term Edgeworth expansion for U statistics of degree two with remainders $o(1/\sqrt{N^*})$ and $O(1/N^*)$. Since we shall often refer to the papers Bloznelis and Götze (1999 a,b,c) we abbreviate them as [BG a,b,c].

A second order asymptotic theory for general asymptotically normal symmetric statistics of *independent and identically distributed* observations was developed in a recent paper by Bentkus, Götze and van Zwet (1997), which concludes a number of previous investigations of particular statistics: Bickel (1974), Callaert and Janssen (1978), Götze (1979), Callaert, Janssen and Veraverbeke (1980), Serfling (1980), Helmers (1982), Helmers and van Zwet (1982), van Zwet (1984), Bickel, Götze and van Zwet (1986), Lai and Wang (1993), etc., This theory is based on the representation of symmetric statistics by sums of U statistics of increasing order via Hoeffding's decomposition. Another approach, see, e.g., Chibisov (1972), Pfanzagl (1973), Bhattacharya and Ghosh (1978), which is based on Taylor expansions of statistics in powers of the underlying i.i.d. observations, focuses on smooth functions of observations.

In view of important classes of applications (jackknife histogram, see, Wu (1990), Shao (1989), Booth and Hall (1993) and subsampling, see, Politis and Romano (1994), Bertail (1997), Bickel, Götze and van Zwet (1997)) we want to develop in this paper a second order asymptotic theory similar to that of Bentkus, Götze and van Zwet (1997) for simple random samples drawn without replacement from finite populations.

The starting point of our asymptotic analysis is the Hoeffding decomposition

$$T = \mathbf{E}T + \sum_{1\le i\le N} g_1(X_i) + \sum_{1\le i<j\le N} g_2(X_i, X_j) + \dots . \tag{1.1}$$

We shall assume that the linear part $\sum g_1(X_i)$ is nondegenerate. That is, $\sigma^2 > 0$, where $\sigma^2 = \mathbf{Var}\, g_1(X_1) > 0$. In the case where, for large N^*, the linear part dominates the statistic we can approximate the distribution of T by a normal distribution, using the central limit theorem. Furthermore, the sum of the linear and quadratic term,

$$U = \mathbf{E}T + \sum_{1 \le i \le N} g_1(X_i) + \sum_{1 \le i < j \le N} g_2(X_i, X_j)$$

typically provides a sufficiently precise approximation to T so that one term Edgeworth expansions for the distribution functions of T and U are the same. Therefore, in order to construct one term Edgeworth expansion of T we do not need to evaluate all the summands of the decomposition (1.1), but (moments of) the first few terms only. An advantage of such an approach is that it provides asymptotic expansions for an *arbitrary* symmetric finite population statistic T no matter whether it is a smooth function of observations or not.

A simple calculation shows that the variance of the linear part satisfies

$$\mathbf{Var} \sum_{i=1}^{N} g_i(X_i) = \sigma^2 \tau^2 \, \frac{n}{n-1}\,, \qquad \tau^2 = npq, \qquad p = N/n, \qquad q = 1 - p.$$

Note that $N^*/2 \le \tau^2 \le N^*$. We shall approximate the distribution function

$$F(x) = \mathbf{P}\{T \le \mathbf{E}T + \sigma\tau x\},$$

by the one term Edgeworth expansion

$$G(x) = \Phi(x) - \frac{(q-p)\,\alpha + 3\,\kappa}{6\,\tau}\,\Phi^{(3)}(x),$$

where

$$\alpha = \sigma^{-3}\mathbf{E}g_1^3(X_1) \qquad \text{and} \qquad \kappa = \sigma^{-3}\tau^2\mathbf{E}g_2(X_1, X_2)g_1(X_1)g_2(X_2).$$

Here $\Phi(x)$ denotes the standard normal distribution function, and $\Phi^{(3)}$ denotes the third derivative of Φ. The expressions for the functions g_1 and g_2 in (1.1) are given by

$$\begin{aligned} g_1(X_i) &= \frac{n-1}{n-N}\,\mathbf{E}(T' \,|\, X_i), \qquad\qquad T' = T - \mathbf{E}T, \\ g_2(X_i, X_j) &= \frac{n-2}{n-N}\,\frac{n-3}{n-N-1}\,\Big(\mathbf{E}(T' \,|\, X_i, X_j) - \frac{n-1}{n-2}\,(\mathbf{E}(T'|X_i) + \mathbf{E}(T'|X_j))\Big), \end{aligned}$$

those for g_k, $k = 3, \ldots, N$, are determined in [BG c].

In order to prove the validity of asymptotic approximations by smooth functions, like $G(x)$, one needs to impose appropriate smoothness conditions on the statistics involved. In the case of linear statistics this is a Cramér type condition, see (1.8) below. Given a general nonlinear statistic T we approximate it by a U-statistic via Hoeffding's decomposition. Therefore, in addition to a Cramér type assumption we need conditions which control the accuracy of such a stochastic expansion, cf. Theorem A below.

2. Smoothness conditions. Introduce the difference operation

$$D^j T = t(X_1, \dots, X_j, \dots, X_N) - t(X_1, \dots, X_j', \dots, X_N), \quad X_j' = X_{N+j},$$

where in the second summand X_j has been replaced by X_j'. In addition, higher order difference operations are defined recursively:

$$D^{j_1,j_2} T = D^{j_2}\big(D^{j_1} T\big), \qquad D^{j_1,j_2,j_3} T = D^{j_3}\big(D^{j_2}(D^{j_1} T)\big), \dots .$$

It is easy to see that the difference operations are symmetric, i.e., $D^{j_1,j_2}T = D^{j_2,j_1}T$, etc. Given $k < N^*$ write

$$\delta_j = \delta_j(T) = \mathbf{E}\big(\tau^{2(j-1)} \mathbb{D}_j T\big)^2, \qquad \mathbb{D}_j T = D^{1,2,\dots,j} T, \qquad 1 \le j \le k.$$

Using the notation

$$U_k(T) = \sum_{1 \le i_1 < \dots < i_k \le N} g_k(X_{i_1}, \dots, X_{i_k})$$

we can write (1.1) as follows

$$T = \mathbf{E}T + U_1(T) + \dots + U_N(T). \tag{1.2}$$

Theorem A. *([BG c]) For $1 \le k < N^*$, we have*

$$T = \mathbf{E}T + U_1(T) + \dots + U_k(T) + R_k, \qquad \text{with} \qquad \mathbf{E}R_k^2 \le \tau^{-2(k-1)} \delta_{k+1}. \tag{1.3}$$

Assume that the population size $n \to \infty$ and the sample size N increases so that $N^* \to \infty$. If σ^2 remains bounded away from zero and

$$\limsup \delta_3 < \infty \tag{1.4}$$

Theorem A implies the relation $T = U + O_P(\tau^{-1})$ thus, showing that up to errors of the second order T/τ and U/τ are asymptotically equivalent.

Recall Cramér's (C) condition for the distribution F_Z of a random variable Z,

$$\sup_{|t|>\delta} |\mathbf{E} \exp\{itZ\}| < 1, \qquad \text{for some} \qquad \delta > 0. \tag{C}$$

Usually, see, e.g., Petrov (1975), this smoothness condition is imposed in addition to the moment conditions in order to establish the validity of asymptotic expansions with remainders $O(N^{-k/2})$ and $o(N^{-k/2})$, $k = 2, 3, \dots$, for the distribution function of the sum of N independent observations from a distribution F_Z. In our situation the condition (C) is too stringent. We shall use a modification of (C) which is applicable to random variables assuming a finite number of values only. For $Z = \sigma^{-1} g_1(X_1)$, we assume that $\rho > 0$, where

$$\rho := 1 - \sup\{ |\mathbf{E} \exp\{itZ\}| : \, b_1/\beta_3 \le |t| \le \tau \}.$$

Here b_1 is a small absolute constant (one may choose, e.g., $b_1 = 0.001$) and

$$\beta_k = \sigma^{-k} \mathbf{E} |g_1(X_1)|^k, \qquad \gamma_k = \sigma^{-k} \tau^{2k} \mathbf{E} |g_2(X_1, X_2)|^k, \qquad k = 2,\, 3,\, 4.$$

Other modifications of Cramér's (C) condition which are applicable to discrete random variables were considered by Albers, Bickel and van Zwet (1976), Robinson (1978) and [BG a], where relations between these conditions are discussed.

3. Results. Write $\zeta = \sigma^{-2} \tau^8 \mathbf{E} g_3^2(X_1, X_2, X_3)$.

Theorem 1.1. *There exists an absolute constant $c > 0$ such that*

$$\Delta := \sup_{x \in R} \bigl| F(x) - G(x) \bigr| \le \frac{c}{\tau^2} \, \frac{\beta_4 + \gamma_4 + \zeta}{\rho^2} + \frac{c}{\tau^2} \, \frac{\delta_4}{\sigma^2 \rho^2}\,. \tag{1.5}$$

For U–statistics of arbitrary but fixed degree k

$$\sum_{1 \le i_1 < \dots < i_k \le N} h(X_{i_1}, \dots, X_{i_k}), \tag{1.6}$$

where h is a real symmetric function defined on k-subsets of $\mathcal{X}$, we have a stronger result.

Theorem 1.2. *There exist an absolute constant $c > 0$ and a constant $c(k) > 0$ depending only on k such that*

$$\Delta \le \frac{c}{\tau^2} \, \frac{\beta_4 + \gamma_4}{\rho^2} + \frac{c(k)}{\tau^2} \, \frac{\delta_3}{\sigma^2 \rho^2}\,. \tag{1.7}$$

Since the absolute constants are not specified Theorems 1.1 and 1.2 should be viewed as asymptotic results.

Assume that the population size $n \to \infty$ and the sample size N increases so that $N^* \to \infty$. In particular, $\tau \to \infty$. In those models where β_4, γ_4 and $\zeta + \delta_4$ (respectively δ_3) remain bounded and

$$\liminf \rho > 0 \tag{1.8}$$

Theorem 1.1 (respectively Theorem 1.2) provides the bound $\Delta = O(\tau^{-2})$. Since $N^*/2 \le \tau^2 \le N^*$ this yields $\Delta = O(1/N^*)$.

Note that this bound is obtained without any additional assumption on p and q. This fact is important for applications, like subsampling, where p or q may tend to zero as $n \to \infty$.

The bound of order $O(\tau^{-2})$ for the remainder is unimprovable, because the next term of the Edgeworth expansion, at least for linear statistics, is of order $O(\tau^{-2})$, see Robinson (1978).

An expansion of the probability $P\{T \le \mathbf{E}T + \sigma\tau x\}$ in powers of τ^{-1} would be the most natural choice of asymptotics. We invoke two simple arguments supporting this choice. Firstly, τ^2 is proportional to the variance of the linear part. Secondly, the number of observations N does not longer determine the scale of T in the case where samples are drawn without replacement since the statistic effectively depends on $N^*(\approx \tau^2)$ observations. Indeed, it was shown in [BG c] that, for $N > n - N$, we have almost surely

$$T = T^*, \qquad T^* = \mathbf{E}T + U_1(T^*) + \cdots + U_{N^*}(T^*), \tag{1.9}$$

where we denote

$$U_k(T^*) = \sum_{1 \le i_1 < \cdots < i_k \le N^*} (-1)^k g_k(X'_{i_1}, \ldots, X'_{i_k}), \qquad X'_j = X_{N+j}.$$

That is, T effectively depends on $N^* = n - N$ observations $X'_1, \ldots, X'_{N^*}$ only.

The bounds of Theorems 1.1 and 1.2 are optimal in the sense that it is impossible to approximate F by a continuous differentiable function, like G, with the remainder $o(\tau^{-2})$, if no additional smoothness condition apart from (1.8) is imposed. Already for U statistics of degree two, Cramér's condition (1.8) together with moment conditions of arbitrary order do not suffice to establish the approximation of order $o(\tau^{-2})$. This fact is demonstrated by means of a counter example in Bentkus, Götze and van Zwet (1997) in the i.i.d. situation, and it is inherited by finite population statistics. See [BG a] for detailed discussions.

Note that the bound (1.5) involves moments (of nonlinear parts) which are higher than those which are necessary to define expansions. Thus, in an optimal dependence on moments one would like to replace $\gamma_4 + \zeta + \delta_4/\sigma^2$ by $\gamma_2 + \delta_3/\sigma^2$ in the remainder.

In order to apply our results to particular classes of statistics one has to estimate moments δ_3 or δ_4 of differences $\mathbb{D}_3 T$ or $\mathbb{D}_4 T$. For U statistics and smooth functions of sample means this problem is easy and routine, see [BG c]. Some applications of our results to resampling procedures are considered in [BG c].

In the case where $n \to \infty$ and N remains fixed the simple random sample model approaches the i.i.d. situation. We have $\tau \to \sqrt{N}$, $p \to 0$, $q \to 1$. Replacing

τ, p and q by $\sqrt{N}$, 0 and 1 respectively we obtain from G the one term Edgeworth expansion for the distribution function of symmetric statistic based on i.i.d. observations, which was constructed in Bentkus, Götze and van Zwet (1997).

The remaining part of the paper is organized as follows. In Section 2 we prove Theorems 1.1 and 1.2. In the proof we use a "data dependent smoothing technique" first introduced in Bentkus, Götze and van Zwet (1997) and expansions of characteristic functions. Expansions of characteristic functions are present separately in Section 3. Section 4 collects auxiliary combinatorial lemmas. Lemma 4.2 of this section may be of independent interest.

2. Proofs

The section consists of two parts. In the first part we collect some facts about Hoeffding's decomposition of finite population statistics. The second part contains proofs of Theorems 1.1 and 1.2.

We shall assume without loss of generality that $\mathbf{E}T = 0$. For $k = 1, 2, \dots,$ we write $\Omega_k = \{1, \dots, k\}$ and $\Omega_k^c = \Omega_N \setminus \Omega_k$.

1. Hoeffding's decomposition (1.2) was studied by Zhao and Chen (1990) in the case of finite population U statistics and by [BG c] in the case of general symmetric finite population statistics.

Given $A = \{i_1, \dots, i_r\} \subset \Omega_n$, with $1 \le r \le N$, and $B = \{j_1, \dots, j_s\} \subset \Omega_n$ write

$$T_A = g_r(X_{i_1}, \dots, X_{i_r}), \qquad \mathbf{E}(T_A|B) = \mathbf{E}(T_A|X_{j_1}, \dots, X_{j_s}),$$

and put $T_\emptyset = 0$. Using this notation we can rewrite (1.2) as follows,

$$T = \sum_{A \in \Omega_N} T_A = U_1(T) + \dots + U_N(T), \quad U_j(T) = \sum_{A \subset \Omega_N,\, |A|=j} T_A, \quad 1 \le j \le N.$$

It is easy to show that in the case of U-statistic of degree k, $2 \le k \le N$, see (1.6), we have $U_j(T) \equiv 0$, for $j \ge k$. An important property of the decomposition is that

$$\mathbf{E}(T_A|B) = 0, \qquad \text{for} \quad A, B \subset \Omega_n, \quad \text{such that} \quad |B| < |A| \le N. \tag{2.1}$$

Note that (2.1) implies $\mathbf{E}U_i(T)U_j(T) = 0$, for $i \ne j$. Therefore, the random variables $U_i(T)$ and $U_j(T)$ are uncorrelated unless $i = j$. For $A, B \subset \Omega_n$ with $1 \le j = |A| = |B| \le N$ and $k = |A \cap B|$, denote

$$\sigma_j^2 = \mathbf{E}T_A^2, \qquad s_{j,k} = \mathbf{E}T_A T_B.$$

Using (2.1) it is easy to show, see e.g., [BG c], that

$$s_{j,k} = (-1)^{j-k} \binom{n-j}{j-k}^{-1} \sigma_j^2, \qquad 0 \le k \le j \le N. \tag{2.2}$$

2. Proofs of Theorems 1.1 and 1.2. We shall assume that $\sigma^2 = \tau^{-2}$.

By $C, c, c_0, c_1, \dots$ we denote positive absolute constants. Given two numbers $a, b > 0$, we write $a \ll b$ if $a \le c\, b$. The expression $\exp\{ix\}$ is abbreviated by $\mathrm{e}\{x\}$. Given a complex function h defined on $\mathbb{R}$, we write $\|H(x)\| = \sup_{x\in\mathbb{R}} |H(x)|$. In the proofs it is more convenient to deal with δ rather than with ρ, where

$$\delta = 1 - \sup\{\mathbf{E}\cos(tg_1(X_1) + s) : \; s \in \mathbb{R}, \; b_1\tau/\beta_3 \le |t| \le \tau^2\}.$$

It is easy to show, see [BG a], that $\rho \le \delta$.

We may and shall assume that for sufficiently small c_0,

$$\beta_4 < c_0\tau^2, \qquad \gamma_2 \le c_0\tau^2\delta^2, \qquad \delta^{-2/3}\ln\tau \le c_0\tau. \tag{2.3}$$

Indeed, if (2.3) fails, the bounds (1.5) and (1.7) follow from the inequalities

$$F(x) \le 1 \qquad \text{and} \qquad |G(x)| \ll 1 + \tau^{-1}(\beta_4^{1/2} + \gamma_2^{1/2}).$$

Note that the first inequality in (2.3) implies that τ is sufficiently large.

Proof of Theorem 1.1. Write $T = \mathcal{U}_3 + R_3$, where $\mathcal{U}_3 = U_1(T) + U_2(T) + U_3(T)$. A Slutzky type argument gives

$$\Delta \le \Delta' + \tau^{-2}\|G^{(1)}(x)\| + \mathbf{P}\{R_3 \ge \tau^{-2}\},$$

where, by (2.3), $\|G^{(1)}(x)\| \ll 1$ and where $\Delta' = \|\mathbf{P}\{\mathcal{U}_3 \le x\} - G(x)\|$ satisfies, by (1.7),

$$\Delta' \ll \tau^{-2}\rho^{-2}(\beta_4 + \gamma_4 + \zeta).$$

Finally, invoking the inequality,

$$\mathbf{P}\{R_3 \ge \tau^{-2}\} \le \delta_4 = \tau^{-2}\delta_4/\sigma^2,$$

see (1.3), we obtain (1.5)

Proof of Theorem 1.2. A typical proof of the validity of an asymptotic expansion for probabilities consists of two main steps: Esseen's smoothing lemma and expansions of characteristic functions. We follow the same line of argument, but instead of the traditional Esseen smoothing lemma we use the "data depending smoothing", procedure introduced by Bentkus, Götze and van Zwet (1997). This a somewhat more sophisticated smoothing technique allows to obtain the optimal rate $O(\tau^{-2})$ for a non-linear statistic, like (1.1), assuming a Cramér type condition on the linear part only.

In view of (1.9) it suffices to prove the theorem in the case where $N \le n/2$. We shall assume that $N \le n/2$. Hence, we have $\tau^2 \le N \le 2\tau^2$ in what follows.

During the proof we skip some technical steps and refer to [BG a], where detailed calculations for these steps are given in the simple case of U-statistics of degree two.

Let m denote the integer closest to the number $8\delta^{-1}\ln\tau$. Due to (2.3), $10 \le m \le N/2$. Split $T = V + W$, where

$$V = \sum_{B\subset\Omega_N,\, B\cap\Omega_m\neq\emptyset} T_B, \qquad W = \sum_{B\subset\Omega_N,\, B\cap\Omega_m=\emptyset} T_B,$$

$$V = \sum_{i=1}^m V_i + \Lambda_m + Y_m + Z_m, \qquad V_i = T_{\{i\}} + \xi_i + \eta_{m,i}, \qquad \xi_i = \sum_{j=m+1}^N T_{\{i,j\}}.$$

Here we denote

$$\Lambda_m = \sum_{B\subset\Omega_m,\, |B|=2} T_B, \qquad Z_m = \sum_{B\subset\Omega_N,\, |B\cap\Omega_m|\ge 3} T_B, \tag{2.4}$$

$$Y_m = \sum_{B\subset\Omega_N,\, |B\cap\Omega_m|=2, |B|\ge 3} T_B, \qquad \eta_{m,i} = \sum_{B\subset\Omega_N,\, B\cap\Omega_m=\{i\},\, |B|\ge 3} T_B.$$

Write $\tilde T = \sum_{i=1}^m V_i + W$ and denote $F_1(x) = \mathbf{P}\{\tilde T \le x\}$. We have $T = \tilde T + R$, where $R = \Lambda_m + Y_m + Z_m$. Given $\varepsilon = \delta^{-2}\tau^{-2}$, a Slutzky type argument gives

$$\Delta \le \Delta_1 + \varepsilon\,\|G^{(1)}(x)\| + \mathbf{P}\{|R| \ge \varepsilon\}, \qquad \Delta_1 = \|F_1(x) - G(x)\|,$$

By Chebyshev's inequality, (4.1) and the inequality $\mathbf{E}|\Lambda_m|^3 \ll m^6\mathbf{E}|T_{\{1,2\}}|^3$,

$$P\{|R| \ge \varepsilon\} \le \mathbf{P}\{|\Lambda_m| \ge \tfrac{\varepsilon}{3}\} + \mathbf{P}\{|Y_m| \ge \tfrac{\varepsilon}{3}\} + \mathbf{P}\{|Z_m| \ge \tfrac{\varepsilon}{3}\} \ll \frac{\delta_3/\sigma^2 + \gamma_3}{\tau^2\delta^2}.$$

Finally, by (2.3), $\|G^{(1)}(x)\| \ll 1$. We obtain $\Delta \ll \Delta_1 + \tau^{-2}\delta^{-2}(1 + \delta_3/\sigma^2 + \gamma_3)$. Therefore, in order to prove (1.7), it remains to bound Δ_1.

Smoothing. Let $\overline{A} = (A_1, \dots, A_n)$ be a random permutation of $(x_1, \dots, x_n)$ uniformly distributed over the class of permutations. Write $r = [(N+m)/2]$ and denote $\mathcal{I}_0 = \{m+1, \dots, N\}$, $\mathcal{J}_0 = \{1, \dots, n\} \setminus \mathcal{I}_0$, $\mathcal{J}_1 = \mathcal{J}_0 \cup \big\{m+1, \dots, r\big\}$ and $\mathcal{J}_2 = \mathcal{J}_0 \cup \big\{r+1, \dots, N\big\}$. Define (random) subpopulations $\mathcal{A}_i = \{A_k,\, k \in \mathcal{J}_i\}$, $i = 0, 1, 2$, and let $\mathcal{A}_i^*$ be a random variable uniformly distributed in $\mathcal{A}_i$.

We assume that $X_j = A_j$, for $j \in \mathcal{I}_0$ and given A_j, $j \in \mathcal{I}_0$, the observations $X_1, \dots, X_m$, are drawn without replacement from $\mathcal{A}_0$. Write

$$f_1(t) = \mathbf{E}\big(\mathrm{e}\{t\tilde T\} \,|\, X_{m+1}, \dots, X_N\big), \qquad \hat F_1(t) = \mathbf{E}\,\mathrm{e}\{t\tilde T\},$$

$$H = N\,\delta/\big(32\,q^{-1}\,N\,(\Theta_1 + \Theta_2) + 1\big), \qquad \Theta_i = \mathbf{E}^*|v_i(A_i^*)|, \quad i = 1, 2,$$

$$v_1(a) = \sum_{r+1\le j\le N} g_2(a, A_j), \qquad v_2(a) = \sum_{m+1\le j\le r} g_2(a, A_j).$$

Here, for $f : \mathcal{X} \to \mathbb{R}$, we denote $\mathbf{E}^* f(A_i^*) = |\mathcal{J}_i|^{-1} \sum_{j \in \mathcal{J}_i} f(A_j)$.

We are going to apply the "data depending smoothing". In order to bound Δ_1 we construct upper bounds for $F_1(x+) - G(x)$ and $G(x) - F_1(x-)$, for every $x \in \mathbb{R}$. Here $F_1(x+)$ (respectively $F_1(x-)$) denotes the right (respectively the left) side limit of F_1 at point x. We have, see Bentkus, Götze and van Zwet (1997),

$$F_1(x+) - G(x) \le \mathbf{E} I_1 + \mathbf{E} I_2 + \mathbf{E} I_3, \tag{2.5}$$
$$I_1 = \frac{1}{2} H^{-1} \int_R \mathrm{e}\{-x\,t\}\, K_1\big(\frac{t}{H}\big) f_1(t)\, dt,$$
$$I_2 = \frac{i}{2\pi} \text{ V.P.} \int_R \mathrm{e}\{-x\,t\} K_2\big(\frac{t}{H}\big)\big(f_1(t) - \hat{G}(t)\big)\,\frac{dt}{t},$$
$$I_3 = \frac{i}{2\pi} \text{ V.P.} \int_R \mathrm{e}\{-x\,t\}\Big(K_2\big(\frac{t}{H}\big) - 1\Big)\hat{G}(t)\,\frac{dt}{t}.$$

Here V.P. denotes Cauchy's principal value, $\hat{G}$ denotes the Fourier-Stieltjes transform of G,

$$K_1(s) = \mathbb{I}\{|s| \le 1\}\,(1 - |s|) \quad \text{and} \quad K_2(s) = \mathbb{I}\{|s| \le 1\}\,((1 - |s|)\,\pi\, s\, \cot(\pi\, s) + |s|).$$

In order to prove an upper bound for $F_1(x+) - G(x)$ we show that

$$|\mathbf{E} I_1| + |\mathbf{E}(I_2 + I_3)| \ll \tau^{-2}\delta^{-2}(\beta_4 + \gamma_4 + c(k)\delta_3/\sigma^2). \tag{2.6}$$

To get the analogous bound for $G(x) - F_1(x-)$ we apply (2.5) to the distribution function of $-\tilde{T}$. In the remaining part of the proof we verify (2.6).

Let us prove (2.6). Introduce the subsets $\mathcal{Z}_i \subset \mathbb{R}$,

$$\mathcal{Z}_1 = \{|t| \le H_1\}, \qquad \mathcal{Z}_2 = \{H_1 \le |t| \le H\}, \qquad \text{where} \quad H_1 = b_1\tau/\beta_3.$$

Obvious decompositions yield, see [BG a],

$$|\mathbf{E}(I_2 + I_3)| \ll |J_1| + \mathbf{E} J_2 + J_3 + R, \quad |\mathbf{E} I_1| \ll |\mathbf{E} J_4| + \mathbf{E} J_5 + R,$$
$$J_1 = \int_{\mathcal{Z}_1} \mathrm{e}\{-tx\}\,\frac{\hat{F}_1(t) - \hat{G}(t)}{t}\,dt, \qquad J_2 = \int_{\mathcal{Z}_2} \frac{|f_1(t)|}{|t|}\,dt,$$
$$J_3 = \int_{|t| > H_1} \frac{|\hat{G}(t)|}{|t|}\,dt, \qquad J_4 = \int_{\mathcal{Z}_1} \mathrm{e}\{-tx\}\,\frac{f_1(t)}{H}\,dt, \qquad J_5 = \int_{\mathcal{Z}_2} \frac{|f_1(t)|}{H}\,dt,$$

with $R = \mathbf{E} H_1^2 H^{-2}$. The bounds $J_3 \ll \tau^{-2}(\beta_4 + \gamma_2)$ and $R \ll \delta^{-2}\tau^{-2}(1 + \gamma_2)$ are proved in [BG a]. Furthermore, note that $J_5 \le J_2$. Therefore, in order to prove (2.6) it suffices to show that

$$\mathbf{E} J_2 \ll \frac{\beta_4 + \delta_3}{\tau^2}, \quad |\mathbf{E} J_4| \ll \frac{1 + \gamma_2 + \delta_3/\sigma^2}{\tau^2\delta^2}, \quad |J_1| \ll \frac{\beta_4 + \gamma_4 + c(k)\delta_3/\sigma^2}{\tau^2\delta^2}. \tag{2.7}$$

The bound for $\mathbf{E}J_2$. Given t denote $\mathbb{I}_t = \mathbb{I}\{|t|\mathbf{E}^*|\eta_m(A_0^*)| < \delta/16\}$, where $\eta_m(x) = \mathbf{E}(\eta_{m,1}|X_1 = x, X_{m+1}, \dots, X_N)$.The identity $f_1(t) = \mathbb{I}_t f_1(t) + (1-\mathbb{I}_t) f_1(t)$ combined with the inequalities

$$1 - \mathbb{I}_t \le 16^2\delta^{-2}t^2\big(\mathbf{E}^*|\eta_m(A_0^*)|\big)^2 \le 16^2\delta^{-2}t^2\mathbf{E}^*\eta_m^2(A_0^*)$$

yields $J_2 \le J_{2,1} + J_{2,2}$, where

$$J_{2.1} = \int_{\mathcal{Z}_2} \mathbb{I}_t \frac{|f_1(t)|}{|t|}\, dt \qquad \text{and} \qquad J_{2.2} = 16^2\delta^{-2}\mathbf{E}^*\eta_m^2(A_0^*)H^2.$$

Invoking the inequality $\mathbf{E}\mathbf{E}^*\eta_m^2(A_0^*) \le N^{-3}\delta_3/\sigma^2$, which follows from (4.1), by symmetry, and using the bound $H \le N\delta$ we obtain $\mathbf{E}J_{2,2} \ll \tau^{-2}\delta_3/\sigma^2$.

In order to bound $\mathbf{E}J_{2,1}$ we proceed as in proof of the inequality (3.12) in [BG a]. The only and minor modification of this proof is related to the new nonlinear term $\eta_{m,i} = \eta_m(X_i)$. Namely, one should replace $v(a) = v_1(a) + v_2(a)$ by $\tilde v(a) = v(a) + \eta_m(a)$ everywhere in the proof of (3.12), ibidem, and use the bound $E^*|t\,\eta_m(A_0^*)| \le \delta/16$ when estimating $\mathbf{E}^*(1+2u_2(A_0^*))$ in (3.17), ibidem. Indeed, this bound holds on the event $\{\mathbb{I}_t \ne 0\}$. The further steps of the proof of (3.12) ibidem given in [BG a] can be adopted without any change and in this way we obtain the bound $\mathbf{E}J_{2.1} \ll \beta_3/N^2$ thus completing the proof of (2.7).

The bound for $\mathbf{E}J_4$. Define J_4' in the same way as J_4, but with $f_1(t)$ replaced by $f_1'(t) = \mathbf{E}(\mathrm{e}\{t\tilde U\}\,|\,\mathcal{I}_0)$, where $\tilde U = U - \Lambda_m$. The identity $\tilde T - \tilde U = R_2 - Y_m - Z_m$, in combination with the inequality $|\,\mathrm{e}\{x\} - \mathrm{e}\{y\}| \le |x-y|$ yields $|\mathbf{E}J_4 - \mathbf{E}J_4'| \ll R$, where

$$\begin{aligned} R &= H_1^2\mathbf{E}H^{-1}\mathbf{E}\big(|R_2 - Y_m - Z_m|\,\big|\,\mathcal{I}_0\big) \\ &\le H_1^2(\mathbf{E}H^{-2})^{1/2}\Big[(\mathbf{E}R_2^2)^{1/2} + (\mathbf{E}Y_m^2)^{1/2} + (\mathbf{E}Z_m^2)^{1/2}\Big], \end{aligned}$$

by Cauchy–Schwartz. It follows from (1.3), (4.1) and the last inequality of (2.3) that $[\dots] \ll \tau^{-2}\delta^{-1}\delta_3^{1/2}/\sigma$. Invoking the inequality $\mathbf{E}H^{-2} \ll N^{-2}\delta^{-2}(1+\gamma_2)$, see (5.1) in [BG a], we obtain $R \ll \tau^{-2}\delta^{-2}(1 + \delta_3/\sigma^2 + \gamma_2)$. Finally, invoking the bound $|\mathbf{E}J_4'| \ll N^{-1}\delta^{-2}(1+\gamma_2)$, see bound for $\mathbf{E}I_6$ in (3.20) ibidem, we obtain (2.7) for $\mathbf{E}J_4$.

The bound for J_1. Given a Borel set $\mathcal{B} \subset \mathbb{R}$ and an integrable complex function f, write $\mathcal{I}_{\mathcal{B}}(f) = \int_{\mathcal{B}} t^{-1}f(t)dt$.

In order to prove the (2.7) for J_1 it suffices to show that

$$|\mathcal{I}_{\mathcal{Z}_1}(\hat F_1 - \hat F)| \ll \tau^{-2}\delta^{-2}(1 + \gamma_2 + \delta_3/\sigma^2), \qquad \hat F(t) = \mathbf{E}\,\mathrm{e}\{tT\}, \tag{2.8}$$

$$|\mathcal{I}_{\mathcal{Z}_1}(\hat F - \hat G)| \ll \tau^{-2}\delta^{-2}(\beta_4 + \gamma_4 + c(k)\delta_3/\sigma^2). \tag{2.9}$$

The proof of (2.9) is given in Section 3. Note that this is the only step in the proof where we use the assumption that T is an U–statistic. It remains to show (2.8). Write $\hat{F}_1(t) = \mathbf{E}\,\mathrm{e}\{t(T-\Lambda_m - Y_m - Z_m)\}$ and expand the exponent in powers of $it(Y_m+Z_m)$ and then in powers of $it\Lambda_m$. We get

$$|\hat{F}_1(t) - \hat{F}(t) - f(t)| \ll \mathbf{E}|tY_m| + \mathbf{E}|tZ_m| + \mathbf{E}t^2\Lambda_m^2, \qquad f(t) = \mathbf{E}\,\mathrm{e}\{tT\}it\Lambda_m.$$

Furthermore, the identity $T - U = R_2$ combined with the mean value theorem yields $|f(t)-g(t)| \ll \mathbf{E}t^2|\Lambda_m R_2|$, where $g(t) = \mathbf{E}\,\mathrm{e}\{tU\}it\Lambda_m$. Therefore,

$$|\mathcal{I}_{\mathcal{Z}_1}(\hat{F}_1 - \hat{F})| \ll |\mathcal{I}_{\mathcal{Z}_1}(g)| + R, \quad R = 2H_1\mathbf{E}|Y_m| + 2H_1|Z_m| + H_1^2\mathbf{E}\Lambda_m^2 + H_1^2\mathbf{E}|\Lambda_m R_2|.$$

Combining Hölder's inequality with (4.1), (1.3) and the inequality $\mathbf{E}\Lambda_m^2 \le m^2 N^{-3}\gamma_2$, see (5.3) in [BG a], we obtain

$$R \ll \tau^{-2}\delta^{-2}(1+\gamma_2+\delta_3/\sigma^2).$$

Finally, invoking the bound

$$|\mathcal{I}_{\mathcal{Z}_1}(g)| \ll \binom{m}{2} N^{-3/2}(1+\gamma_2) \ll N^{-1}\delta^{-2}(1+\gamma_2),$$

see the bound for I_{15}, below (3.38) ibidem, we obtain (2.8) thus, completing the proof of the theorem.

3. Expansions

In order to prove (2.9) we shall show that

$$\text{(3.1)} \qquad \mathcal{I}_{\mathcal{B}_1}(\hat{F} - \hat{G}) \ll \mathcal{R}, \qquad \mathcal{I}_{\mathcal{B}_2}(\hat{F} - \hat{G}) \ll \mathcal{R}, \qquad \mathcal{R} = \frac{\beta_4 + \gamma_4 + c(k)\delta_3/\sigma^2}{\tau^2\delta^2},$$

where $\mathcal{B}_1 = \{|t| \le c_1\}$ and $\mathcal{B}_2 = \{c_1 \le |t| \le H_1\}$ and where the constant c_1 will be specified later.

The first inequality of (3.1) follows from the bounds

$$\text{(3.2)} \quad \mathcal{I}_{\mathcal{B}_1}(\hat{F} - \hat{F}_U) \ll \mathcal{R} \quad \text{and} \quad \mathcal{I}_{\mathcal{B}_1}(\hat{F}_U - \hat{G}) \ll \mathcal{R}, \qquad \hat{F}_U(t) = \mathbf{E}\,\mathrm{e}\{tU\},$$

where the first bound of (3.2) follows from the inequalities $|\hat{F}(t) - \hat{F}_U(t)| \le \mathbf{E}|tR_2|$ and $\mathbf{E}|R_2| \le (\mathbf{E}R_2^2)^{1/2}$ and (1.3). The second bound of (3.2) is proved in [BG a], formula (4.1)

In order to prove the second inequality of (3.1) we write the characteristic function $\hat{F}(t)$ in Erdős-Rényi (1959) form, see (3.4) below. Let $\nu = (\nu_1, \dots, \nu_n)$ be i.i.d.

Bernoulli random variables independent of $(X_1, \dots, X_n)$ and having probabilities $\mathbf{P}\{\nu_1 = 1\} = p$ and $\mathbf{P}\{\nu_1 = 0\} = q$. Observe, that the conditional distribution of

$$T^* = \sum_{A \subset \Omega_n} T_A \nu_A^*, \qquad \text{where} \qquad \nu_A^* = \prod_{i \in A} \nu_i,$$

given the event $\mathcal{E} = \{S_\nu = N\}$, where $S_\nu = \nu_1 + \dots + \nu_n$, coincides with the distribution of T. Therefore, $\hat{F}$ can be written as follows

$$\hat{F}(t) = \lambda \int_{-\pi\tau}^{\pi\tau} \mathbf{E}\, \mathrm{e}\{tT^* + \tau^{-1} s(S_\nu - N)\} ds, \qquad \lambda^{-1} = 2\pi \mathbf{P}\{\mathcal{E}\}\tau. \tag{3.3}$$

Using (2.1) it is easy to show that, for $1 \le k \le N$, almost surely

$$\sum_{A \subset \Omega_n,\, |A| = k} T_A \nu_A^* = \sum_{A \subset \Omega_n,\, |A| = k} Q_A, \qquad Q_A = T_A \tilde{\nu}_A, \qquad \tilde{\nu}_A = \prod_{i \in A} (\nu_i - p).$$

Therefore, almost surely, $tT^* + \tau^{-1} s(S_\nu - N) = S + tQ$, where

$$S = \sum_{i=1}^n S_i, \quad S_i = (tT_{\{i\}} + \tau^{-1}s)(\nu_i - p), \qquad Q = \sum_{A \subset \Omega_n,\, 2 \le |A| \le N} Q_A.$$

Substituting this identity in (3.3), we obtain

$$\hat{F}(t) = \lambda \int_{-\pi\tau}^{\pi\tau} \mathbf{E}\, \mathrm{e}\{S + tQ\} ds. \tag{3.4}$$

In view of (3.4), the second inequality of (3.1) follows from the inequalities

$$\int_{t \in \mathcal{B}_2} \lambda \int_{|s| \le \pi\tau} \big| \mathbf{E}\, \mathrm{e}\{S + tQ\} - (h_1 + h_2) \big| ds \, \frac{dt}{|t|} \ll \mathcal{R}, \tag{3.5}$$

$$\int_{t \in \mathcal{B}_2} \Big| \lambda \int_{|s| \le \pi\tau} (h_1 + h_2) ds - \hat{G}(t) \Big| \, \frac{dt}{|t|} \ll \mathcal{R}, \tag{3.6}$$

$$h_1 = \mathbf{E}\, \mathrm{e}\{S\}, \qquad h_2 = i^3 \binom{n}{2} \mathbf{E}\, \mathrm{e}\{S_3 + \dots + S_n\} V, \qquad V = tQ_{\{1,2\}} S_1 S_2.$$

The inequality (3.6) is obtained by combining the proof of the analogous bound in the i.i.d. situation, see Lemma 6.1 of Bentkus, Götze and van Zwet(1997), with the proof of the Berry–Esseen bound for finite population sample means given in Höglund (1978), see also [BG a].

It remains to prove (3.5). Let $f(s,t)$ and $g(s,t)$ denote two complex functions. In order to reduce the notations we write $f \prec \mathcal{R}$ if

$$\int_{\mathcal{B}_2} \frac{dt}{|t|} \int_{-\pi\tau}^{\pi\tau} |f(s,t)| ds \ll \mathcal{R}$$

and write $f \sim g$ if $f - g \prec \mathcal{R}$. In view of the inequality $\lambda \le \sqrt{2}\pi$, see Höglund (1978), the bound (3.5) is equivalent to the relation $\mathbf{E}\, \mathrm{e}\{S + tQ\} \sim h_1 + h_2$.

Let us prove (3.5). In what follows assume $t \in \mathcal{B}_2$ and $|s| \le \pi\tau$. Given s, t write $u = s^2 + t^2$ and let m be the integer closest to the number $c_2 n u^{-1} \ln u$, where the constant c_2 will be specified later. We shall choose c_1 and c_2 so that $10 < m < n/2$. Split

$$\text{(3.7)} \quad Q = K + L + W + Y + Z, \qquad K = \zeta + \mu, \quad \zeta = \sum_{j=1}^{m} \zeta_j, \quad \mu = \sum_{j=1}^{m} \mu_j,$$

$$\zeta_j = \sum_{A \cap \Omega_m = \{j\},\, |A|=2} Q_A, \qquad \mu_j = \sum_{A \cap \Omega_m = \{j\},\, |A| \ge 3} Q_A, \qquad 1 \le j \le m,$$

$$L = \sum_{A \subset \Omega_m,\, |A|=2} Q_A, \qquad Y = \sum_{|A \cap \Omega_m|=2,\, |A| \ge 3} Q_A,$$

$$Z = \sum_{|A \cap \Omega_m| \ge 3} Q_A, \qquad W = \sum_{A \cap \Omega_m = \emptyset,\, |A| \ge 2} Q_A,$$

and denote $f_1 = \mathbf{E}\, \mathrm{e}\{S + t(K+W)\}$ and $f_2 = \mathbf{E}\, \mathrm{e}\{S + t(K+W)\} itL$.

In order to prove $\mathbf{E} \exp\{S + tQ\} \sim h_1 + h_2$ we shall show that

$$\text{(3.8)} \qquad \mathbf{E}\, \mathrm{e}\{S + tQ\} \sim f_1 + f_2,$$

$$\text{(3.9)} \qquad f_2 \sim h_3, \qquad h_3 = i^3 \binom{m}{2} \mathbf{E}\, \mathrm{e}\{S_3 + \cdots + S_n\} V,$$

$$\text{(3.10)} \qquad f_1 \sim h_1 + h_2 - h_3.$$

We first introduce some notation. Given a sum $v = v_1 + \cdots + v_k$ we denote $v_B = \sum_{j \in B} v_j$, for $B \subset \Omega_k$. Given $B \subset \Omega_n$ we write $\mathbf{E}_{(B)}$ to denote the conditional expectation given all the random variables, but ν_j, $j \in B$. For $D \subset \Omega_m$, denote

$$\mathbb{Y}_D = |\mathbf{E}_{(D)}\, \mathrm{e}\{S_D\}|, \quad \mathbb{Z}_D = |\mathbf{E}_{(D)}\, \mathrm{e}\{S_D + t\zeta_D\}|, \quad \mathbb{I}_D = \mathbb{I}\{\varkappa_D > c_3^{-1}\},$$

where $\varkappa_D = \tau^2 |D|^{-1} \sum_{j \in D} \zeta^2(X_j)$, with $\zeta(a) = \sum_{j=m+1}^{n} g_2(a, X_j)(\nu_j - p)$, satisfies

$$\text{(3.11)} \quad \mathbf{E}(\varkappa_D \,|X_i, X_j) \ll \tau^{-2}\gamma_2, \qquad \mathbf{E}(\varkappa_D^{3/2} \,|X_i, X_j) \ll \tau^{-3}\gamma_3, \quad \forall\, i, j \in \Omega_m \setminus D.$$

Proceeding as in proof of Lemma 5.4 in [BG a], see also formulas (4.9), (4.10) ibidem, for every $D \subset \Omega_m$, with cardinality $|D| \geq m/4$, we can construct a random variable $\mathbb{F}_D$ depending only on X_i, $i \in D$, such that

$$(3.12)\quad \mathbb{Y}_D \leq \mathbb{F}_D, \quad \mathbb{Z}_D \leq \mathbb{I}_D + \mathbb{F}_D, \qquad \mathbf{E}(\mathbb{F}_D^2 \,|X_i,\, X_j) \leq c\, u^{-20}, \quad \forall\, i,\, j \in \Omega \setminus D,$$

almost surely. In this step, i.e. in the proof of (3.12), the constants c_1, c_2 and c_3 are specified, see [BG a]. Here we only mention that c_1 and c_2 are choosen so that the inequality $10 \leq m \leq n/2$ holds as well.

Split $\Omega_m = \Omega_m^1 \cup \Omega_m^2 \cup \Omega_m^3$, where Ω_m^i, $i = 1, 2, 3$ are disjoint consecutive intervals with cardinalities $|\Omega_m^i| \approx m/3$. For $i \leq j$, let $\Omega_{i,j}$ denote the set of all pairs $\{l, r\}$ such that $l \in \Omega_m^i$, $r \in \Omega_m^j$ and $l < r$.

Proof of (3.8). Expanding the exponent in powers of itZ and invoking (3.19) we get $\mathbf{E}\,\mathrm{e}\{S + itQ\} = f_3 + R$, where $f_3 = \mathbf{E}\,\mathrm{e}\{S + t(K + L + W + Y)\}$ and where

$$|R| \leq \mathbf{E}|tZ| \leq |t|(\mathbf{E}Z^2)^{1/2} \ll |t| c_1^{1/2}(k) u^{3/2} \ln^{3/2} u\, \delta_3^{1/2} \sigma^{-1} \tau^{-2} \prec \mathcal{R}.$$

Furthermore, expanding in powers of $it(L + Y)$ we obtain

$$f_3 = f_1 + f_2 + f_4 + R, \qquad f_4 = \mathbf{E}\,\mathrm{e}\{S + t(K + W)\} itY,$$
$$|R| \ll t^2 \mathbf{E}(L + Y)^2 \ll t^2 u^{-2} \ln^2 u\, \tau^{-2}(\gamma_2 + c_1(k)\delta_3/\sigma^2) \prec \mathcal{R}.$$

In the last step we invoked (3.20) and used the identity $\mathbf{E}L^2 = \binom{m}{2} p^2 q^2 \tau^{-6} \gamma_2$. We obtain $\mathbf{E}\,\mathrm{e}\{S + tQ\} \sim f_1 + f_2 + f_4$.

It remains to prove that $f_4 \prec \mathcal{R}$. To this aim we show that $f_4 \sim f_5$ and $f_5 \prec \mathcal{R}$, where $f_5 = \mathbf{E}\,\mathrm{e}\{S + t(\zeta + W)\} itY$. By the mean value theorem $|f_4 - f_5| \leq \mathbf{E}t^2|Y\mu|$. Furthermore, by Cauchy–Schwartz and (3.20), (3.21),

$$\mathbf{E}t^2|Y\mu| \leq t^2(\mathbf{E}Y^2)^{1/2}(\mathbf{E}\mu^2)^{1/2} \ll t^2 c_1(k) u^{-3/2} \ln^{3/2} u\, \tau^{-4} \delta_3/\sigma^2 \prec \mathcal{R}.$$

Therefore, $f_4 \sim f_5$. In order to prove $f_5 \prec \mathcal{R}$ we split $f_5 = \sum_{1 \leq i \leq j \leq 3} f_{i,j}$ and show that $f_{i,j} \prec \mathcal{R}$, for every $i \leq j$. Here $f_{i,j}$ are defined in the same way as f_5, but with Y replaced by $Y_{i,j}$, with $Y_{i,j}$ denoting the sum of all Q_A such that $A \cap \Omega_m \in \Omega_{i,j}$. Given i, j choose r from $\Omega_3 \setminus \{i, j\}$ and note that the random variable $Y_{i,j}$ and the sequence $\{\nu_l,\, l \in \Omega_m^r\}$ are independent. Therefore, by (3.12),

$$|f_{i,j}| \leq |t| \mathbf{E}\mathbb{Z}_{\Omega_m^r} |Y_{i,j}| \leq |t|(\mathbf{E}\mathbb{Z}_{\Omega_m^r}^2)^{1/2} (\mathbf{E}Y_{i,j}^2)^{1/2} \leq |t|(\mathbf{E}\mathbb{F}_{\Omega_m^r}^2 + \mathbf{E}\varkappa_{\Omega_m^r})^{1/2} (\mathbf{E}Y_{i,j}^2)^{1/2}.$$

Note that the bound (3.20) applies to $\mathbf{E}Y_{i,j}^2$ as well. This bound in combination with (3.12) and (3.11) implies $f_{i,j} \prec \mathcal{R}$ thus completing the proof of (3.8).

Proof of (3.9). Write $f_6 = \mathbf{E}\,\mathrm{e}\{S + t(\zeta + W_0)\}itL$, where W_0 denotes the sum of all Q_A such that $A \cap \Omega_m = \emptyset$ and $|A| = 2$. It is shown in formula (4.15) of [BG a] that $f_6 \sim h_3$.

Let us prove $f_2 \sim f_6$. By the mean value theorem, $|f_2 - f_7| \le t^2 \mathbf{E}|L\mu|$, where $f_7 = \mathbf{E}\,\mathrm{e}\{S + t(\zeta + W)\}itL$. Invoking (3.21) and the bound $\mathbf{E}L^2 \le m^2p^2q^2\tau^{-6}\gamma_2$ we obtain $t^2\mathbf{E}|L\mu| \prec \mathcal{R}$, by Cauchy–Schwartz. Hence, $f_2 \sim f_7$.

It remains to show $f_7 \sim f_6$. Split

$$f_6 = \sum_{1\le i\le j\le 3} f'_{i,j} \qquad \text{and} \qquad f_7 = \sum_{1\le i\le j\le 3} f^*_{i,j},$$

were $f'_{i,j}$ (respectively $f^*_{i,j}$) is defined in the same way as f_6 (respectively f_7), but with L replaced by $L_{i,j} = \sum_{A\in\Omega_{i,j}} Q_A$. It suffices to prove $f'_{i,j} \sim f^*_{i,j}$, for every $i \le j$. Given $i \le j$, choose $r \in \Omega_3 \setminus \{i, j\}$ and write

$$f^*_{i,j} = \mathbf{E}\,\mathrm{e}\{S + t(\zeta + W_0 + W_1)\}itL_{i,j}, \qquad W_1 = \sum_{A\cap\Omega_m=\emptyset,\, |A|\ge 3} Q_A. \tag{3.13}$$

Expanding the exponent in powers of itW_1 we obtain $f^*_{i,j} = f'_{i,j} + t^2R$, where

$$|R| \le \mathbf{E}\mathbb{Z}_{\Omega^r_m}|L_{i,j}W_1| \le R_1 + R_2, \quad R_1 = \mathbf{E}\mathbb{F}_{\Omega^r_m}|L_{i,j}W_1|, \quad R_2 = \mathbf{E}\mathbb{I}_{\Omega^r_m}|L_{i,j}W_1|,$$

by (3.12). Furthermore, by Cauchy–Schwartz,

$$R_1^2 \le \mathbf{E}W_1^2\mathbf{E}L^2_{i,j}\mathbb{F}^2_{\Omega^r_m}, \qquad R_2^2 \le \mathbf{E}W_1^2\mathbf{E}L^2_{i,j}\mathbb{I}_{\Omega^r_m} \le \mathbf{E}W_1^2\mathbf{E}L^2_{i,j}\varkappa_{\Omega^r_m} \tag{3.14}$$

Fix $\{i_1, i_2\} \in \Omega_{i,j}$. By symmetry, (3.12) and (3.11),

$$\mathbf{E}L^2_{i,j}\mathbb{F}^2_{\Omega^r_m} = |\Omega_{i,j}|p^2q^2\mathbf{E}g_2^2(X_{i_1}, X_{i_2})\mathbf{E}(\mathbb{F}^2_{\Omega^r_m}\,|i_1, i_2) \ll \tau^{-2}u^{-20}\gamma_2, \tag{3.15}$$

$$\mathbf{E}L^2_{i,j}\varkappa_{\Omega^r_m} = |\Omega_{i,j}|p^2q^2\mathbf{E}g_2^2(X_{i_1}, X_{i_2})\mathbf{E}(\varkappa_{\Omega^r_m}\,|i_1, i_2) \ll u^{-2}\ln^2 u\,\tau^{-4}\gamma_2^2. \tag{3.16}$$

Here we estimated $|\Omega_{i,j}| < m^2$ and $m^2p^2q^2 \ll \tau^4u^{-2}\ln^2 u \ll \tau^4$. It follows from (3.14), (3.15), (3.16) and (3.22) that $t^2R \prec \mathcal{R}$. We obtain $f'_{i,j} \sim f^*_{i,j}$ thus proving $f_7 \sim f_6$.

Proof of (3.10). Expanding in powers of $it\mu$ we obtain

$$f_1 = f_9 + f_{10} + R, \qquad f_9 = \mathbf{E}\,\mathrm{e}\{S + t(\zeta + W)\}, \quad f_{10} = \mathbf{E}\,\mathrm{e}\{S + t(\zeta + W)\}it\mu,$$

where $|R| \le t^2\mu^2 \prec \mathcal{R}$, by (3.21).

Let us show $f_{10} \prec \mathcal{R}$. We now split $\mu = \mu^*_1 + \mu^*_2 + \mu^*_3$, where $\mu^*_j = \mu_{\Omega^j_m}$. It suffices to prove $f^*_j = \mathbf{E}\,\mathrm{e}\{S + t(\zeta + W)\}it\mu^*_j \prec \mathcal{R}$, for $1 \le j \le 3$. Given j, fix $r \in \Omega_3 \setminus \{j\}$. By (3.12) and Cauchy–Schwartz,

$$|f^*_j| \le |t|\mathbf{E}\mathbb{Z}_{\Omega^r_m}|\mu^*_j| \le |t|(\mathbf{E}(\mu^*_j)^2)^{1/2}(\mathbf{E}\mathbb{F}^2_{\Omega^r_m} + \mathbf{E}\varkappa^{3/2}_{\Omega^r_m})^{1/2}.$$

Note that (3.21) holds for μ_j^* as well. This bound in combination with (3.11) and the last inequality of (3.12) gives $f_j^* \prec \mathcal{R}$. Hence, we obtain $f_2 \sim f_9$. Furthermore, proceeding as in proof of (4.35) in [BG a], we get

$$(3.17) \qquad f_9 \sim f_{11} + f_{12}, \qquad f_{11} = \mathbf{E}\, \mathrm{e}\{S + tW\}, \qquad f_{12} = \mathbf{E}\, \mathrm{e}\{S + tW\} it\zeta.$$

In the next step we shall show that

$$(3.18) \quad f_{11} \sim f_{13}, \quad f_{12} \sim f_{14}, \quad f_{13} = \mathbf{E}\, \mathrm{e}\{S + tW_0\}, \quad f_{14} = \mathbf{E}\, \mathrm{e}\{S + tW_0\} it\zeta.$$

Recall that $W = W_0 + W_1$, where W_0 and W_1 are defined in the proof of (3.9) above. It follows from (3.17) and (3.18) that $f_1 \sim f_{13} + f_{14}$. Invoking the relation $f_{13} + f_{14} \sim h_1 + h_2 - h_3$, see (4.36), (4.37) ibidem, we obtain (3.10). We complete the proof of (3.10) by verifying (3.18).

Expanding in powers of itW_1 we get $f_{11} = f_{13} + R$, where $|R| \le \mathbf{E}\mathbb{Y}_{\Omega_m}|tW_1| \prec \mathcal{R}$, by Cauchy–Schwartz, (3.12) and (3.22). In order to prove $f_{12} \sim f_{14}$ split $\Omega_m = V_1 \cup V_2$, where $V_1 \cap V_2 = \emptyset$ with cardinality $|V_i| \approx m/2$, for $i = 1, 2$, and write $\zeta = \zeta_{V_1} + \zeta_{V_2}$. Expanding in powers of itW_1, we get $f_{12} = f_{14} + R_{(1)} + R_{(2)}$, where $|R_{(i)}| \le \mathbf{E}t^2|W_1\zeta_{V_i}|\mathbb{Y}_{\Omega_m \setminus V_i}$. Fix $r \in V_1$. By Cauchy–Schwartz and symmetry,

$$\begin{aligned} |R_{(1)}|^2 &\le t^2 \mathbf{E}W_1^2 \mathbf{E}\zeta_{V_1}^2 \mathbb{Y}_{V_2}^2 = t^2 \mathbf{E}W_1^2 |V_1|(n-m)p^2q^2 \mathbf{E}g_2^2(X_r, X_n)\mathbf{E}(\mathbb{Y}_{V_2}^2 | X_r, X_n) \\ &\ll t^2 \tau^{-6} u^{-20}, \end{aligned}$$

by (3.22) and (3.12). We have $R_{(1)} \prec \mathcal{R}$. The same bound holds for $R_{(2)}$ as well. The proof of (3.10) is complete.

Lemma 3.1. *Assume that T is a U–statistic of degree k. Then*

$$(3.19) \qquad \mathbf{E}Z^2 \le c_1(k) m^3 p^3 q^3 \tau^{-10} \delta_3 \sigma^{-2},$$

$$(3.20) \qquad \mathbf{E}Y^2 \le c_1(k) m^2 p^2 q^2 \tau^{-8} \delta_3 / \sigma^2,$$

$$(3.21) \qquad \mathbf{E}\mu^2 \le c_1(k) mpq \tau^{-6} \delta_3 / \sigma^2,$$

$$(3.22) \qquad \mathbf{E}W_1^2 \le \tau^{-4} \delta_3 / \sigma^2,$$

where Y, Z and μ are defined in (3.7) and W_1 is defined in (3.13).

Proof of Lemma 3.1. Given a random variable $Q_{\mathcal{H}} = \sum_{A \in \mathcal{H}} Q_A$, where $\mathcal{H}$ is some class of subsets $\mathcal{H} = \{A \subset \Omega_n,\ 2 \le |A| \le N\}$, write

$$\hat{e}_j(Q_{\mathcal{H}}) = \sigma_j^{-2}(pq)^{-j}\ \mathbf{Var}\Big(\sum_{A \in \mathcal{H},\, |A| = j} Q_A \Big),$$

if $\sigma_j^2 > 0$ and put $\hat{e}_j(Q_{\mathcal{H}}) = 0$ otherwise. Since Q_A and Q_B are uncorrelated unless $A = B$, we have $\mathbf{E}Q_{\mathcal{H}}^2 = \sum_{j=2}^N \hat{e}_j(Q_{\mathcal{H}})p^j q^j \sigma_j^2$. Therefore, in order to prove $\mathbf{E}Q_{\mathcal{H}}^2 < c(\mathcal{H})\tau^{-10}\delta_3/\sigma^2$ it suffices to show that

$$(3.23)\quad \hat{e}_j(Q_{\mathcal{H}})p^j q^j \le c(\mathcal{H})e_j(Z_3), \qquad e_j(Z_3) = \binom{N-3}{j-3}\binom{n-N}{j-3}\binom{n-j}{j-3}^{-1},$$

and make use of (A.23). Note that for U–statistics of degree k we have $j \le k$.

Let us prove (3.19). A simple calculation shows that, for $m \le n/2$,

$$\hat{e}_j(Z) = \sum_{k=0}^{j-3} \binom{m}{j-k}\binom{n-m}{k} \le m^3 n^{j-3} \le c_1(k) m^3 p^{3-j} q^{3-j} e_j(Z_3),$$

where the last inequality holds for $3 \le j \le k$, with some constant $c_1(k)$. In view of (3.23), we obtain (3.19).

The proof of (3.20), (3.21) and (3.22) is almost the same. We have

$$\hat{e}_j(Y) = \binom{m}{2}\binom{n-m}{j-2}, \quad \hat{e}_j(\mu) = m\binom{n-m}{j-1}, \quad \hat{e}_j(W_1) = \binom{n-m}{j},$$

and therefore,

$$\hat{e}_j(Y) \le m^2 n^{j-2}, \qquad \hat{e}_j(\mu) \le m n^{j-1}, \qquad \hat{e}_j(W_1) \le n^j.$$

Finally, invoking the inequality $m^r n^{j-r} p^j q^j \le c_1(k) m^r p^r q^r \tau^{2(3-r)} e_j(Z_3)$, for $r = 0, 1, 2$, we obtain (3.22), (3.21), (3.20).

4. Combinatorial lemmas

Here we construct bounds for the second moments of the random variables defined in (2.4) above.

Lemma 4.1. *Assume that* $100 \le N \le n/2$. *Given* $3 \le m \le N$ *and* $i \in \Omega_m$, *we have*

$$(4.1)\qquad \mathbf{E}Y_m^2 \ll N^{-3} m^4 \delta_3, \qquad \mathbf{E}Z_m^2 \ll N^{-4} m^6 \delta_3, \qquad \mathbf{E}\eta_{m,i}^2 \ll N^{-2}\delta_3.$$

As a by-product of the proof of Lemma 4.1 we obtain a formula for the sums (4.2), see Lemma 4.2 below, which might be of independent interest.

Let us first introduce some notation. Write $\Omega_k^c = \Omega_N \setminus \Omega_k$. Introduce the following random variables

$$S = \sum_{A\in\mathcal{H}} T_A = U_1(S) + \cdots + U_N(S), \qquad U_j(S) = \sum_{A\in\mathcal{H},\, |A|=j} T_A, \quad 1 \le j \le N,$$

where $\mathcal{H}$ denotes a class of subsets $A \subset \Omega_n$, with $|A| \le N$. Observe, that, by (2.1), the random variables $U_i(S)$ and $U_j(S)$ are uncorrelated unless $i = j$. In particular, we have

$$\mathbf{E}S^2 = \sum_{1\le j\le N} \mathbf{E}U_j^2(S) = \sum_{1\le j\le N} e_j(S)\sigma_j^2, \qquad e_j(S) := \mathbf{E}U_j^2(S)\sigma_j^{-2}.$$

If $\sigma_j^2 = 0$, we put $e_j(S) = 0$. For non-negative integers k, s, t, u denote

$$r_k(s,t,u) = \sum_{v=0}^{s\wedge t} (-1)^{v+k} \binom{s}{v}\binom{t}{v}\binom{u}{v+k}^{-1}, \qquad u \ge s\wedge t + k, \tag{4.2}$$

and put $r_k(s,t,u) = 0$, for $u < s \wedge t + k$. Recall that $s \wedge t = \min\{s,t\}$ and, for $x \in \mathbb{R}$, $\binom{x}{r} = [x]_r/r!$, if the integer $r \ge 0$, and $\binom{x}{r} = 0$, for $r < 0$. Here $[x]_r = x(x-1)\cdots(x-r+1)$, for $r > 0$, and $[x]_0 = 1$. In particular, for non-negative integers $s < v$, we have $\binom{s}{v} = 0$

Proof of Lemma 4.1. The proof consists of two steps, (4.3) and (4.4),

$$\mathbf{E}Z_m^2 \ll m^6 \mathbf{E}Z_3^2, \qquad \mathbf{E}Y_m^2 \ll N\, m^4 \mathbf{E}Z_3^2, \qquad \mathbf{E}\eta_{m,i}^2 \ll N^2 \mathbf{E}Z_3^2, \tag{4.3}$$

$$\mathbf{E}Z_3^2 \ll N^{-4}\delta_3, \qquad \text{for} \qquad N \ge 100. \tag{4.4}$$

Proof of (4.3). Let us show that, for $3 \le j \le N$,

$$e_j(Z_3) = \binom{N-3}{j-3} r_0(j-3, N-j, n-j) = \binom{N-3}{j-3}\binom{n-N}{j-3}\binom{n-j}{j-3}^{-1}. \tag{4.5}$$

By symmetry,

$$\mathbf{E}U_j^2(Z_3) = \binom{N-3}{j-3}\mathbf{E}T_{\Omega_j}U_j(Z_3), \quad \mathbf{E}T_{\Omega_j}U_j(Z_3) = \sum_{v=0}^{(j-3)\wedge(N-j)} M_v s_{j,j-v},$$

where $M_v = \binom{N-j}{v}\binom{j-3}{v}$ counts the summands T_A of the sum $U_j(Z_3)$ satisfying $|A \cap \Omega_j| = j - v$. Invoking (2.2), we obtain $\mathbf{E}T_{\Omega_j}U_j(Z_3) = r_0(j-3, N-j, n-j)\sigma_j^2$, thus proving the first part of (4.5). For the second part use (4.24).

Let us prove the first inequality of (4.3). Write $Z_m = Z_3 + D_4 + \cdots + D_m$, where $D_k = Z_k - Z_{k-1}$. By the inequality $(a_1 + \cdots + a_k)^2 \le k(a_1^2 + \cdots + a_k^2)$,

$$\mathbf{E}Z_m^2 \le (m-2)(\mathbf{E}Z_3^2 + \mathbf{E}D_4^2 + \cdots + \mathbf{E}D_m^2).$$

Now (4.3) follows from the inequalities $\mathbf{E}D_k^2 \ll k^4 \mathbf{E}Z_3^2$, $4 \le k \le m$. To prove these inequalities we show that, for $4 \le k \le m$ and $3 \le j \le N$,

$$e_j(D_k) \ll k^4 e_j(Z_3), \tag{4.6}$$

Observe, that $e_j(D_k) = 0$, for $N - k + 3 < j \le N$, by the definition of D_k. In the case where $3 \le j \le N - k + 3$, the inequalities (4.6) follow from the relations (4.8), (4.9) and (4.11), which we shall prove below.

We have

$$D_k = \sum_{A \subset \Omega_{k-1},\, |A|=2} \; \sum_{B \subset \Omega_k^c} T_{A \cup \{k\} \cup B}. \tag{4.7}$$

Given j, the sum (4.7) has $\binom{k-1}{2}\binom{N-k}{j-3}$ different summands $T_{A\cup\{k\}\cup B}$, such that $|A \cup \{k\} \cup B| = j$. Fix $B_0 \subset \Omega_k^c$ with $|B_0| = j - 3$ and denote $A_0 = \Omega_2 \cup \{k\} \cup B_0$. By symmetry

$$\mathbf{E}U_j^2(D_k) = \binom{k-1}{2}\binom{N-k}{j-3}\mathbf{E}T_{A_0}U_j(D_k). \tag{4.8}$$

In the next step we show that

$$\mathbf{E}T_{A_0}U_j(D_k) = \Bigl(L_0(j) + 2(k-3)L_1(j) + \binom{k-3}{2}L_2(j)\Bigr)\sigma_j^2, \tag{4.9}$$

where we denote $L_i(j) = r_i(j-3, \kappa, n-j)$, where $\kappa = N - k - j + 3$. To this aim split $U_j(D_k) = W_0 + W_1 + W_2$, where

$$W_i = \sum_{A \in \mathcal{A}_i} \; \sum_{B \subset \Omega_k^c,\, |B|=j-3} T_{A\cup\{k\}\cup B}, \quad \mathcal{A}_i = \{A \subset \Omega_{k-1} : |A| = 2,\ |A \cap \Omega_2| = 2 - i\},$$

and write $\mathbf{E}T_{A_0}U_j(D_k) = \mathbf{E}T_{A_0}W_0 + \mathbf{E}T_{A_0}W_1 + \mathbf{E}T_{A_0}W_2$. Note that (4.9) follows form the identities

$$\mathbf{E}T_{A_0}W_i = |\mathcal{A}_i| L_i(j)\sigma_j^2, \quad i = 0, 1, 2, \tag{4.10}$$

which we are going to prove. Denote $H_0 = \{1, 2\}$, $H_1 = \{1, 3\}$, $H_2 = \{3, 4\}$. By symmetry, for $i = 0, 1, 2$,

$$\mathbf{E}T_{A_0}W_i = |\mathcal{A}_i| \mathbf{E}T_{A_0} \sum_{B \subset \Omega_k^c,\, |B|=j-3} T_{H_i\cup\{k\}\cup B} = |\mathcal{A}_i| \sum_{v=0}^{(j-3)\wedge\kappa} M_v\, s_{j,j-v-i}.$$

Here $M_v = \binom{\kappa}{v}\binom{j-3}{v}$ counts those subsets $B \subset \Omega_k^c$ such that $|B \cap B_0| = j - 3 - v$. Invoking (2.2) we obtain (4.10).

We complete the proof of (4.6), by showing that, for $3 \le j \le N - k + 3$,

$$L'_i(j) := \binom{N-k}{j-3}|L_i(j)|/e_j(Z_3) \ll 1, \qquad i = 0, 1, 2. \tag{4.11}$$

To evaluate the ratio $L'_i(j)$ we use the expression (4.5) for $e_j(Z_3)$ and invoke the formulas (4.26). Simple calculations yield

$$L'_0(j) = \frac{[N-k]_{j-3}}{[n-N]_{j-3}} \; \frac{[n-N+k-3]_{j-3}}{[N-3]_{j-3}} \; = \prod_{r=0}^{j-4} \frac{x_r}{y_r} \; \frac{y_r + k - 3}{x_r + k - 3}\,,$$

where we denote $x_r = N - k - r$ and $y_r = n - N - r$. Now $L'_0(j) \le 1$ follows from the inequalities $x_r \le y_r$, which are consequences of the inequality $N \le n/2$.

The proof of (4.11) for $i = 1, 2$ is similar, but somewhat more involved, because the expression (4.26) for $L_i(j)$ becomes more complex, for $i = 1, 2$.

Let us prove the second inequality of (4.3). To this aim we shall show that

$$e_j(Y_m) \ll N\, m^4 e_j(Z_3), \qquad 3 \le j \le N. \tag{4.12}$$

Note that $e_j(Y_m) = 0$, for $j > N - m + 2$, by the definition of Y_m. Let us prove (4.12), for $3 \le j \le N - m + 2$. Given j, fix $B_0 \subset \Omega_m^c$, with $|B_0| = j - 2$. By symmetry,

$$\mathbf{E}U_j^2(Y_m) = \binom{m}{2}\binom{N-m}{j-2}\mathbf{E}T_{\Omega_2 \cup B_0} U_j(Y_m). \tag{4.13}$$

Proceeding as in the proof of (4.9) above, we obtain

$$\mathbf{E}T_{\Omega_2 \cup B_0} U_j(Y_m) = \Big(L_0(j) + 2(m-2)L_1(j) + \binom{m-2}{2} L_2(j)\Big)\sigma_j^2, \tag{4.14}$$

where $L_i(j) = r_i(j-2, N-m-j+2, n-j)$, for $i = 0, 1, 2$. Furthermore, arguing as in proof of (4.11) we show that, for $3 \le j \le N - m + 2$,

$$\binom{N-m}{j-2} L_i(j) \ll N\, e_j(Z_3), \quad \text{for} \quad i = 0, 1, 2. \tag{4.15}$$

Now (4.12) follows from (4.13), (4.14) and (4.15).

In order to prove the last inequality of (4.3) we shall show that, for $3 \le j \le N$,

$$e_j(\eta_{m,i}) \le N^2 e_j(Z_3). \tag{4.16}$$

Note that $e_j(\eta_{m,i})=0$, for $j>N-m+1$, by the definition of $\eta_{m,i}$. Let us prove (4.16), for $3\le j\le N-m+1$. Given j, fix $B_0\subset\Omega_m^c$, with $|B_0|=j-1$. By symmetry,

$$\mathbf{E}U_j^2(\eta_{m,i})=\binom{N-m}{j-1}\mathbf{E}T_{\{i\}\cup B_0}U_j(\eta_{m,i}). \tag{4.17}$$

Denote $\kappa=N-m-j+1$. A direct calculation shows that

$$\mathbf{E}T_{\{i\}\cup B_0}U_j(\eta_{m,i})=\sum_{v=0}^{(j-1)\wedge\kappa}\binom{j-1}{v}\binom{\kappa}{v}s_{j,j-v}=r_0(j-1,\kappa,n-j)\sigma_j^2,$$

where in the last step we invoke (2.2). Furthermore, proceeding as in the proof of (4.11) we obtain

$$\binom{N-m}{j-1}r_0(j-1,\kappa,n-j)\le N^2e_j(Z_3). \tag{4.18}$$

Now (4.16) follows from (4.17) and (4.18).

Proof of (4.4). Note that the inequality $N\le n/2$ implies $\tau^2\ge N/2$. Therefore, in order to prove (4.4) it suffices to show that $\mathbf{E}Z_3^2\ll\mathbf{E}(\mathbb{D}_3T)^2$. For this purpose we show that

$$e_j(Z_3)\ll e_j(\mathbb{D}_3T),\qquad 3\le j\le N. \tag{4.19}$$

Let us evaluate $e_j(\mathbb{D}_3T)$. Denoting $L_k=r_k(j-3,N-j,n-j)$ we have

$$e_j(\mathbb{D}_3T)=\binom{N-3}{j-3}(8L_0-24L_1+24L_2-8L_3). \tag{4.20}$$

To verify (4.20) write $\Omega_{3,0}=\Omega_3$ and $\Omega_{3,k}=\{1',\dots,k'\}\cup\{k+1,\dots,3\}$, for $k=1,2,3$, where $i'=N+i$. By symmetry,

$$\begin{aligned}&\mathbf{E}U_j^2(\mathbb{D}_3T)=8M_0-24M_1+24M_2-8M_3,\\ &M_k=\mathbf{E}V_0V_k,\quad V_k=\sum_{B\subset\Omega_3^c,\,|B|=j-3}T_{\Omega_{3,k}\cup B}.\end{aligned} \tag{4.21}$$

Again, by symmetry, we get $M_k=\binom{N-3}{j-3}\mathbf{E}T_{\Omega_j}V_k$. Invoking (2.2) we obtain

$$\mathbf{E}T_{\Omega_j}V_k=\sum_{v=0}^{(j-3)\wedge(N-j)}\binom{j-3}{v}\binom{N-j}{v}s_{j,j-v-k}=L_k\sigma_j^2.$$

Therefore, $M_k = \binom{N-3}{j-3} L_k \sigma_j^2$ and now (4.20) follows from (4.21).

In view of (4.5) and (4.20) we can write

$$e_j(\mathbb{D}_3 T) = 8(1-W)e_j(Z_3), \qquad W = (3L_1 - 3L_2 + L_3)/L_0. \tag{4.22}$$

Invoking the formulas (4.26) for L_k, we find that

$$W < W_1 + W_2, \qquad W_1 = 3A'_{1,1} - 3A'_{2,2} + A'_{3,3}, \quad W_2 = 12A'_{2,1} + 18A'_{3,1},$$

where $A'_{k,r} = A_{k,r}/L_0$ and where the coefficients $A_{k,r}$ are given by (4.26). A simple analysis shows that $W_1 \le c_4 < 1$ and $W_2 \ll N^{-1}$. Therefore, for large N (calculations show that $N > 100$ suffice) we have $W \le c_5 < 1$, for $3 \le j \le N$. Now (4.19) follows from (4.22).

In the remaining part of the section we evaluate the sum

$$r_k(s,t,u) = \sum_{v=0}^{s\wedge t} (-1)^{v+k} \binom{s}{v} \binom{t}{v} \binom{u}{v+k}^{-1}.$$

Using the identity, see Feller (1968) Chapter II,

$$\sum_v (-1)^v \binom{a}{v} \binom{u-v}{t} = \binom{u-a}{u-t}, \qquad a,t,u \in \{0,1,2,\dots\}. \tag{4.23}$$

Zhao and Chen (1990) showed that

$$r_0(s,t,u) = \binom{u-s}{t} \binom{u}{t}^{-1}, \qquad u \ge s \wedge t. \tag{4.24}$$

Given $k = 0,1,2,\dots$, let $l(k,r)$, $0 \le r \le k$, denote the coefficients of the expansion

$$[x+k]_k = l(k,k)[x]_k + l(k,k-1)[x]_{k-1} + \cdots + l(k,0)[x]_0, \qquad x \in R. \tag{4.25}$$

Lemma 4.2. *Let $k,s,t,u \in \{0,1,2,\dots\}$. For $u \ge s \wedge t + k$, we have*

$$r_k(s,t,u) = \sum_{r=0}^{k} l(k,r)(-1)^{r+k} A_{k,r}, \tag{4.26}$$

where $A_{k,r} = \dfrac{(u-s-k)!(u-t-k)![s]_r[t]_r}{(u-s-t-k+r)!u!}$, *for $u \ge s+t+k-r$, and $A_{k,r} = 0$ otherwise.*

Clearly, the numbers $l(i,j)$ can be expressed by Stirling numbers. A direct calculation shows that

$$l(0,0)=1, \quad l(1,0)=1, \quad l(1,1)=1, \qquad l(2,0)=2, \quad l(2,1)=4, \quad l(2,2)=1,$$
$$l(3,0)=6, \quad l(3,1)=18, \quad l(3,2)=9, \quad l(3,3)=1.$$

Proof of Lemma 5.1. Write $a = s \wedge t$ and $b = s \vee t$. We have

$$r_k(s,t,u) = \sum_{v=0}^{a} (-1)^{v+k} \binom{b}{v} M_v, \quad \text{where} \quad M_v = \binom{a}{v} \binom{u}{v+k}^{-1}. \tag{4.27}$$

A simple calculation shows that

$$M_v = [v+k]_k \binom{u-k-v}{u-k-a} w_k(a,u), \qquad w_k(a,u) = \binom{u-k}{a}^{-1} [u]_k^{-1}.$$

Invoking the expansion (4.25), for the function $v \to [v+k]_k$, we obtain an expression for M_v. Substituting this expression in (4.27) we get

$$r_k(s,t,u) = w_k(a,u) \sum_{r=0}^{k} l(k,r) S_r, \qquad S_r = \sum_{v=0}^{a} (-1)^{v+k} \binom{b}{v} \binom{u-k-v}{u-k-a} [v]_r.$$

We complete the proof of (4.26) by showing that, for $0 \le r \le k$,

$$S_r = (-1)^{r+k} [b]_r \binom{u-b-k}{a-r} \quad \text{and} \quad [b]_r \binom{u-b-k}{a-r} w_k(a,u) = A_{k,r}. \tag{4.28}$$

Note that $[v]_r = 0$, for $v < r$. For $r \le v \le b$, we have $[v]_r \binom{b}{v} = [b]_r \binom{b-r}{v-r}$. Therefore, denoting $v' = v - r$, we can write

$$S_r = \sum_{v'=0}^{a-r} (-1)^{v'+r+k} [b]_r \binom{b-r}{v'} \binom{u-k-r-v'}{u-k-a}.$$

Finally, invoking (4.23) we obtain the first identity of (4.28). The second identity is trivial.

References

Albers, W., Bickel, P.J. and van Zwet, W.R., *Asymptotic expansions for the power of distribution free tests in the one-sample problem*, Ann. Statist. **4** (1976), 108–156.

Babu, Gutti Jogesh, Bai, Z. D., *Mixtures of global and local Edgeworth expansions and their applications*, J. Multivariate Anal. **59** (1996), 282–307.

Babu, G.J. and Singh, K., *Edgeworth expansions for sampling without replacement from finite populations*, J. Multivar. Analysis. **17** (1985), 261–278.

Bentkus, V., Götze, F. and van Zwet, W. R., *An Edgeworth expansion for symmetric statistics,*, Ann. Statist. **25** (1997), 851–896.

Bertail, P., *Second-order properties of an extrapolated bootstrap without replacement under weak assumptions*, Bernoulli **3** (1997), 149–179.

Bhattacharya, R.N. and Ghosh, J.K., *On the validity of the formal Edgeworth expansion*, Ann. Statist. **6** (1978), 434–451.

Bickel, P.J., *Edgeworth expansions in non parametric statistics*, Ann. Statist. **2** (1974), 1–20.

Bickel, P.J. and Robinson, J., *Edgeworth expansions and smoothness*, Ann. Probab. **10** (1982), 500–503.

Bickel, P.J., Götze, F. and van Zwet, W. R., *The Edgeworth expansion for U-statistics of degree two*, Ann. Statist. **14** (1986), 1463–1484.

Bickel, P.J. and van Zwet, W. R., *Asymptotic expansions for the power of distribution free tests in the two-sample problem*, Ann. Statist. **6** (1978), 937–1004.

Bikelis, A., *On the estimation of the remainder term in the central limit theorem for samples from finite populations.*, Studia Sci. Math. Hungar. **4** (1972), 345–354. (In Russian).

Bloznelis, M., *A Berry–Esseen bound for finite population Student's statistic*, Ann. Probab. (1999), To Appear.

Bloznelis, M. and Götze, F., *An Edgeworth expansion for finite population U-statistics*, Bernoulli (1999 a), To Appear.

Bloznelis, M. and Götze, F., *One term Edgeworth expansion for finite population U–statistics of degree two*, Acta Applicandae Mathematicae **58** (1999 b), 1–16.

Bloznelis, M. and Götze, F., *Orthogonal decomposition of finite population statistic and its applications to distributional asymptotics*, Preprint 99 – 143, SFB 343 Universität Bielefeld (1999 c).

Bolthausen, E. and Götze, F., *The rate of convergence for multivariate sampling statistics*, Ann. Statist. **21** (1993), 1692–1710.

Booth, J. G. and Hall, P., *An improvement of the jackknife distribution function estimator*, Ann. Statist. **21** (1993), 1476–1485.

Callaert, H., Janssen, P., *A Berry–Esseen theorm for U–statistics*, Ann. Statist. **6** (1978), 417–421.

Callaert, H., Janssen, P. and Veraverbeke, N., *An Edgeworth expansion for U-statistics*, Ann. Statist. **8** (1980), 299–312.

Chibisov, D.M., *Asymptotic expansion for the distribution of a statistic admitting a stochastic expansion*, Theory Probab. Appl. **17** (1972), 620–630.

Erdös, P. and Rényi, A., *On the central limit theorem for samples from a finite population*, Publ. Math. Inst. Hungar. Acad. Sci. **4** (1959), 49–61.

Feller, W., *An introduction to probability theory and its applications. I*, Wiley, New York, 1968.

Götze, F., *Asymptotic expansions for bivariate von Mises functionals*, Z. Wahrsch. verw. Gebiete **50** (1979), 333–355.

Götze, F. and van Zwet, W.R., *Edgeworth expansions for asymptotically linear statistics* (1991), Preprint 91 - 034, SFB 343, Universität Bielefeld.

Helmers, R., *Edgeworth expansions for linear combinations of order statistics* **105** (1982), Math. Centre Tracts, Math. Centrum, Amsterdam.

Helmers, R., and van Zwet, W. R., *The Berry–Esseen bound for U-statistics.* **1** (1982), Statistical Decision Theory and Related Topics,III. Vol. 1. (S.S. Gupta and J.O. Berger, eds.), Academic Press, New York, 497–512.

Höglund, T., *Sampling from a finite population: a remainder term estimate*, Scand. J. Stat. **5** (1978), 69–71.

Kokic, P.N. and Weber, N.C., *An Edgeworth expansion for U-statistics based on samples from finite populations*, Ann. Probab. **18** (1990), 390–404.

Lai, T. L. and Wang, J.Q., *Edgeworth expansions for symmetric statistics with applications to bootstrap methods*, Statist. Sinica **3** (1993), 517–542.

Nandi, H. K. and Sen, P.K., *On the properties of U-statistics when the observations are not independent.II. Unbiased estimation of the parameters of a finite population.*, Calcutta Statist. Assoc. Bull. **12** (1963), 124–148.

Petrov, V.V., *Sums of independent random variables*, Springer, New York (1975).

Pfanzagl, J., *Asymptotic expansions related to minimum contrast estimators*, Ann. Statist. **1** (1973), 993–1026.

Prawitz, H., *Limits for a distribution, if the characteristic function is given in a finite domain*, Scand. AktuarTidskr. (1972), 138–154.

Politis, D. N. and Romano, J. P., *Large sample confidence regions based on subsamples under minimal assumptions*, Ann. Statist. **22** (1994), 2031–2050.

Rao, C. R., Zhao, L. C., *Berry-Esseen bounds for finite-population t-statistics*, Statist. Probab. Lett. **21** (1994), 409–416.

Robinson, J., *An asymptotic expansion for samples from a finite population*, Ann. Statist. **6** (1978), 1005–1011.

Schneller, W., *Edgeworth expansions for linear rank statistics*, Ann. Statist. **17** (1989), 1103–1123.

Serfling, R. J., *Approximation Theorems of Mathematical Statistics*, Wiley, New York, 1980.

Shao, J., *The efficiency and consistency of approximations to the jackknife variance estimators*, J. Amer. Statist. Assoc. **84** (1989), 114–119.

Wu, C.F.J., *On the asymptotic properties of the jackknife histogram*, Ann. Statist. **18**, 1438–1452.

Zhao, L. C. and Chen, X. R., *Berry–Esseen bounds for finite-population U-statistics*, Sci.Sinica Ser. A **30** (1987), 113–127.

Zhao, L. C. and Chen, X. R., *Normal approximation for finite-population U-statistics*, Acta Math. Appl. Sinica **6** (1990), 263–272.

van Zwet, W.R., *A Berry-Esseen bound for symmetric statistics*, Z. Wahrsch. verw. Gebiete **66** (1984), 425–440.

Mindaugas Bloznelis
Dept. of Mathematics and Informatics
Vilnius University
Naugarduko 24
Vilnius 2006
Lithuania
E-mail address: `Mindaugas.Bloznelis@maf.vu.lt`

Friedrich Götze
Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld 1
Germany
E-mail address: `goetze@mathematik.uni-bielefeld.de`