São Paulo Journal of Mathematical Sciences **8**, 2 (2014), 285–309

Quasi U-statistics of infinite order and applications to the subgroup decomposition of some diversity measures

Aluísio Pinheiro

Department of Statistics, University of Campinas

Pranab Kumar Sen

Department of Biostatistics, UNC at Chapel Hill

Abstract. In several applications, information is drawn from qualitative variables. In such cases, measures of central tendency and dispersion may be highly inappropriate. Variability for categorical data can be correctly quantified by the so-called diversity measures. These measures can be modified to quantify heterogeneity between groups (or subpopulations). Pinheiro et al. (2005) shows that Hamming distance can be employed in such way and the resulting estimator of heterogeneity between populations will be asymptotically normal under mild regularity conditions.

Pinheiro et al. (2009) proposes a class of weighted U-statistics based on degenerate kernels of degree 2, called quasi U-statistics, with the property of asymptotic normality under suitable conditions. This is generalized to kernels of degree m by Pinheiro et al. (2011). In this work we generalize this class to an infinite order degenerate kernel. We then use this powerful tools and the reverse martingale nature of U-statistics to study the asymptotic behavior of a collection of transformed classic diversity measures. We are able to estimate them in a common framework instead of the usual individualized estimation procedures.

MSC 2000: primary - 62G10; secondary - 62G20, 92D20.

1. Introduction

Measures of diversity have been extensively studied in the past century. Seminal works are motivated either by economics (Gini, 1912), genetics

Key words: within-populations diversity measures, between -populations diversity measures, asymptotic normality, U-statistics, non-standard asymptotics.

Acknowledgment of support: this research was funded by FAPESP (13/00506-1, 13/16952-0) and CNPq 304512/2011-7.

(Simpson, 1949), information theory (Shannon, 1948) or ecology (Williams, 1945), among other fields.

Rao (1982c) formulates some characterization theorems that relates the following diversity measures: the Gini-Simnpson index; the Shannon index; the α-order entropy of Havrda and Charavát; the paired Shannon entropy; the α-degree entropy of Renyi; and the γ-entropy function. Salicrú et al. (2005) studies the hypothesis of homogeneity of within-populations diversity measures under a general framework.

Those works have established mathematically and statistically the fundamental properties of those diversity mesures: they can be employed to provide a reseacher with estimates and tests for within-populations dissimilarity measures. Between-population measures include the Mahalanobis distance (Mahalanobis, 1936) and Nei's distance (Nei, 1972;Nei and Roychoudhury, 1974; Nei, 1978).

The aforementioned measures of diversity are defined for measuring one characteristic at a time: a single locus in genetic studies; a single variable in economics or ecology; a single channel in communications, etc. Areas of application for diversity measures include: phylogenetis (Anselmo and Pinheiro, 2012; Moulton et al., 2007; Dress and Steel, 2007); population genetics (Gillet, 2007; Gilbert et al., 2005, Kussell and Leibler, 2005); time series analysis (Valk and Pinheiro, 2012); and economics (Nayak and Gastwirth, 1989).

It is interesting to generalize the univariate indices for the employment as multidimensional indices of diversity. Particularly, for genomic data sets, the number of characteristics is usually very large, each one being gene expression or DNA pair of bases and so forth. Moreover, dependencies between loci or genes is a well-known fact (Tavaré and Giddings, 1989; Gillooly et al., 2005; Politi et al., 2005; Pinheiro et al., 2006; Kim et al., 2008).

From a parametric or semi-parametric point of view careful study of the underlying dependence structure should be considered. With such knowledge, multidimensional diversity measures could then generalize the one-dimensional statistics. We approach this generalization from a purely non-parametric paradigm (Sen, 1999; Pinheiro et al., 2005). Other than robustness against ill-posed assumptions, this procedure will provide a subgroup decomposability which allows us to compute within and between-population diversity based upon then same original diversity measures.

Frees (1989) studies the properties of infinite order U-statistics. It proves asymptotic normality for nondegenerate kernels under a $\sqrt{n}$ growth on the kernel approximation. Pinheiro et al. (2005) shows that a suitable contrast of U-statistics naturally arising in genomic studies is asymptotically normal

even though each U-statistics kernel is degenerate. Pinheiro et al. (2011) proposes a class of degree m quasi U-statistics (and variants) for which asymptotic normality can be proven under suitable conditions.

We generalize these results to infinite order quasi U-statistics and then utilize these powerful tools to study a class of diversity measures, equipping it with a common estimation procedure. This procedure will attain asymptotic normality under mild conditions.

The text goes as follows. In Section 2 we present the general idea of decomposability of symmetric statistics in terms of the decomposability of its finite-degree components. In Section 3, some classical diversity results are presented, as well as their multi-dimensional representations. In Section 4 we introduce the class of quasi U-statistics with infinite order and some of its most important properties, including the asymptotic normality of its elements. The special case of Hamming distance-like measures is detailed in Section 5. Discussion and some final remarks are presented in Section 6.

2. The General Idea of Decomposability of Symmetric Statistics

Let $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ be i.i.d random vectors of dimension K (which may increase) with common distribution F such that

$$F \equiv \lambda_1 F_1 + \cdots + \lambda_G F_G, \tag{1}$$

where $F_1, \ldots, F_G$ are probability distribution functions, $G \geq 2$ and $\lambda_1 + \cdots + \lambda_G = 1$.

Consider G groups defined by F_g, $g = 1, \ldots, G$. The samples are defined by random vectors $\mathbf{Y}1^{(g)}, \mathbf{Y}_2^{(g)}, \ldots$ i.i.d. F_g, $g = 1, \ldots, G$. Let $\theta \equiv \theta(\boldsymbol{\lambda}, F)$, $\theta^{(g)} = \theta(1_g, F_g)$, $g = 1, \ldots, G$ (θ_g is the projection of θ on F_g) , and suppose $\theta = E_F \phi(\mathbf{Y}_1, \ldots)$ for a concave function ϕ, i.e.,

$$\theta \geq \lambda_1 \theta^{(1)} + \cdots + \lambda_G \theta^{(G)}, \tag{2}$$

with equality if and only if $F \equiv F_1 \equiv \cdots \equiv F_G$ and/or $\max_g \lambda_g = 1$.

Define a sequence of functionals θ_m, $m \geq 2$, so that

$$\theta = \lim_{m \to \infty} \theta_m. \tag{3}$$

Let

$$\begin{aligned} \theta_m^{(g)} &= E\phi_m(\mathbf{Y}_1^{(g)}, \ldots, \mathbf{Y}_m^{(g)}),\ g = 1, \ldots, G \\ \theta_m^{(g_1, \ldots, g_m)} &= E\phi_m(\mathbf{Y}_1^{(g_1)}, \ldots, \mathbf{Y}_m^{(g_m)}), \end{aligned}$$

where the latter is taken for m_g random vectors from F_g so that $0 \leq m_g \leq m$, $g = 1, \ldots, G$, and $m_1 + \cdots + m_G = m$. By (3), $\theta_m^{(g)} \to \theta^{(g)}$ for each

$g = 1, \ldots, G$. Choose $\lambda_g = m_g/m$, for $g = 1, \ldots, G$. One has by (2)

$$\theta_m^{(g_1,\ldots,g_m)} \geq m_1\theta_m^{(1)} + \cdots + m_G\theta_m^{(G)},$$

with equality iff $\theta_m^{(1)} \equiv \cdots \equiv \theta_m^{(G)}$ and/or $\max_g m_g = m$.

We can then interpret the G random samples as a single sample from F of size $n = \sum_{g=1}^{G} n_g$ with mixture probabilities m_g/m, $g = 1, \ldots, G$. Each $(\lambda_1, \cdots, \lambda_G)$ is an element of the (G-1)-simplex $\mathcal{S}_{G-1}$. Consequently, any $(m_1, \cdots, m_G)$ is such that $(m_1/m, \cdots, m_G/m)$ is an element of $\mathcal{S}_{G-1} \bigcap \mathcal{L}_m^G$, where $\mathcal{L}_m$ is the lattice formed by real numbers between 0 and 1 with grid steps $1/m$.

3. **Some Classical Diversity Measures**

Some metrics have been widely used for the analysis of qualitative data. They include the *Hamming*, *Nei* and *Mahalanobis distances*. Based on these metrics, statistical measurement of variability can be pursued. The aim of the analysis may be measuring variability for homogeneous populations (within variability) or between populations (between variability). Initially, the diversity measures were built for individual characteristics or specific locus but they can be generalized to incorporate multivariate elements.

The *Gini-Simpson index* of diversity (Gini, 1912; Simpson, 1949; Nei, 1972; Lewontin, 1972) is defined as follows. Let $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_C\}$ be the probabilities of C alleles (nucleotide or amino acid) at a locus in a population. The gene diversity at that locus is

$$h = 1 - \sum_{i=1}^{C} \pi_i^2 \quad . \tag{1}$$

One may estimate h by the plug-in estimator $\widehat{h}$, but Nei & Roychoudhury (1974) and Nei (1978) present other parametric alternatives which can outperform the Gini-Simpson plug-in estimator.

Another popular choice of diversity measure is the *Shannon Information Index*

$$h_s = -\sum_{i=1}^{C} \pi_i \log_e \pi_i, \tag{2}$$

motivated in information theory by entropy concepts. It has been successfully applied in ecology and evolutionary genetics (Lewontin, 1972; Rao, 1982a,b; Magurran, 1988). Its plug-in estimator is also biased (Hutcheson, 1970; Bowman et al., 1971), although there are simple parametric alternatives for bias correction (Peet, 1974).

The index

$$h_H(\boldsymbol{\pi}) = \left[\sum_{i=1}^{C} \pi_i^{\alpha}\right]^{1/(1-\alpha)} \tag{3}$$

is the basis of another approach to unify diversity measures (Hill, 1973).

Rao (1982c) formulates a characterization theorem which states that, if a measure of diversity $h(\boldsymbol{\pi}) = h(\pi_1, \ldots, \pi_C)$ satisfies

(i) $h(\boldsymbol{\pi})$ is symmetric with respect to the components of $\boldsymbol{\pi}$ and attains its maximum when all C categories are equally frequent
(ii) $h(\boldsymbol{\pi})$ admits partial derivatives up to the second order of the $C-1$ independent components of $\boldsymbol{\pi}$ and the matrix of second partial derivatives, $h"(\boldsymbol{\pi}) = (h"_{ij}(\boldsymbol{\pi}))$ for $i, j = 1, 2, \ldots, C-1$ with $h"_{ij}(\boldsymbol{\pi}) = \partial^2 h(\boldsymbol{\pi})/\partial\pi_i\partial\pi_j$, is continuous and not null at $\boldsymbol{\pi} = \mathbf{e} \equiv (1/C, \ldots, 1/C)$
(iii) $h\{(\boldsymbol{\pi} + \mathbf{e})/2\} = \frac{1}{2}\{h(\boldsymbol{\pi}) + h(\mathbf{e})\} = k\{h(\mathbf{e}) - h(\boldsymbol{\pi})\}$, where k is a constant,

then $h(\boldsymbol{\pi})$ must be of the form

$$h(\boldsymbol{\pi}) = a\left[1 - \sum_{i=1}^{C} \pi_i^2\right] + b \tag{4}$$

where $a > 0$ and b are constants, i.e. it characterizes the Gini-Simpson gene diversity index (1) and diversity measures based on it.

On the other hand, the Shannon-Information Index given by (2) and the four indices, the *α-order entropy of Havrda and Charavát*, given by

$$h_\alpha(\boldsymbol{\pi}) = [1 - \sum_{i=1}^{C} \pi_i^{\alpha}]/[2^{\alpha-1} - 1] \text{ for } \alpha > 0 \quad \text{and } \alpha \neq 1,$$

the *paired Shannon entropy*, given by

$$h_p(\boldsymbol{\pi}) = -\sum_{i=1}^{C} \pi_i \ln \pi_i - \sum_{i=1}^{C} (1 - p_i) \ln(1 - \pi_i),$$

the *α-degree entropy of Renyi*, given by

$$h_R(\boldsymbol{\pi}) = (1-\alpha)^{-1} \ln(\sum_{i=1}^{C} \pi_i^{\alpha}) \quad \text{for } 0 < \alpha < 1,$$

and the γ-*entropy function*, given by

$$h_\gamma(\boldsymbol{\pi}) = [1 - (\sum_{i=1}^{C} \pi_i^{1/\gamma})^\gamma]/[1 - 2^{\gamma-1}] \quad \text{for } \gamma > 0,\ \gamma \neq 1,$$

satisfy the following two conditions

C_1: $h(\boldsymbol{\pi}) = 0$ if and only if all components of $\boldsymbol{\pi}$ are zero except for one (i.e., $\pi_i = 1$ for one i and the remaining π_i's are all zero)

C_2: $h\{\lambda\,\boldsymbol{\pi} + (1-\lambda)\boldsymbol{\theta}\} \geq \lambda h(\boldsymbol{\pi}) + (1-\lambda)h(\boldsymbol{\theta})$, with equality if and only if $\boldsymbol{\pi} = \boldsymbol{\theta}$ (concavity property),

as seen in Rao (1982a) and Rao & Boudreau (1984).

Salicru et al. (2005) proposes homogeneity tests for such class of within-populations univariate diversity measures. In many applications, including DNA and amino-acid sequences, one would like to study several characteristics. That can be done by suitable generalizations of the above measures of diversity. Pinheiro et al. (2011) presents a class of test statistics for a general class of multivariate diversity measures which take as special cases diversity measures of finite order. In this work we go a step further incorporating infinite order diversity measures to the class of asymptotically normal test statistics defined in Pinheiro et al. (2011).

Consider K characteristics (or loci) and C categories. Let π_{ck} the respective probability for the c-th category on the k-th characteristic, $k = 1, \ldots, K$, and $c = 1, \ldots, C$.

One can see that the generalized Hamming metric is given by

$$h_{Ha}(\pi) = 1 - \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}^2,$$

the generalized Shannon can be written as

$$h_{Sh}(\pi) = -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}\ln\pi_{ck},$$

the generalized α-th order entropy of Havrda and Charavát, as

$$h_{HC,\alpha} = \left[1 - \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}^{\alpha}\right] / \left[2^{\alpha-1} - 1\right]$$

for $0 < \alpha \neq 1$, the generalized paired Shannon entropy, as

$$h_{pSh}(\pi) = -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}\ln\pi_{ck} - \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}(1-\pi_{ck})\ln(1-\pi_{ck}),$$

the generalized α-degree entropy of Renyi, as

$$h_{R,\alpha}(\pi) = (1-\alpha)^{-1} \ln \left(\frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \pi_{ck}^{\alpha} \right)$$

for $0 < \alpha < 1$, and the γ-entropy function, as

$$h_{E,\gamma} = \left[1 - \left(\frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \pi_{ck}^{1/\gamma} \right)^{\gamma} \right] / \left[1 - 2^{\gamma-1} \right],$$

for $0 < \gamma \neq 1$.

These generalizations due to the length of the sequences do not preclude us from further generalizations through conveniently weighing in characteristics and/or categories. For the sake of notational simplicity we proceed on the ordinary form above but the results hold for the general weighted diversity measures as well.

4. A Class of Infinite Order Quasi U-statistics with a Martingale Representation and Asymptotic Normality

The results in this section generalizes to an infinite order kernel the class introduced in Pinheiro et al. (2011). Some of the proofs are shorter, and we direct the readers to this previous manuscript for further details. Define T_n as the following function of a symmetric kernel ϕ:

$$T_n = \sum_{i_1,\ldots,i_{m_n}}^{1,n} \eta_{n,i_1\cdots i_{m_n}} \phi(\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_{m_n}}). \tag{1}$$

We assume the following conditions:

(i) the degree of the kernel is increasing as the sample size increases, i.e. $m_n \to \infty$, but $m_n = o(n)$ as $n \to \infty$,

(ii) $\eta_{n,i_1\cdots i_{m_n}}$ are such that

$$\sum_{i_1,\ldots,i_{m_n}}^{1,n} \eta_{n,i_1\cdots i_{m_n}} = 0, \tag{2}$$

and

$$\sum_{i_1,\ldots,i_{m_n}}^{1,n} \eta_{n,i_1\cdots i_{m_n}}^2 = M_n(\text{which increases in } n \geq m_n), \tag{3}$$

(iii) $\sum_{i_1,\ldots,i_{m_n}}^{1,n}$ is taken on all strictly ordered permutations of $1, \ldots, n$,

(iv) $\phi(\cdot,\ldots,\cdot)$ is a kernel of degree m_n, stationary of order r_n $(1 \leq r_n < m_n)$, for which we let $\theta_{m_n} = \mathrm{E}\phi(\mathbf{Y}_1,\ldots,\mathbf{Y}_{m_n})$, and

(v) $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ are i.i.d. random vectors of dimension K.

The class of weighted U-statistics defined by (1), which we call quasi U-statistics (Pinheiro et al., 2011) possess a natural martingale representation. From this, one is able to ascertain asymptotic normality via a martingale central limit theorem. For the martingale property, one assumes that $\phi(\cdot,\ldots,\cdot)$ is a symmetric stationary kernel of order $r_n = m_n - 1$, centered at 0, and forms an orthogonal system for which

$$\mathrm{E}[\phi(\mathbf{Y}_1,\ldots,\mathbf{Y}_{m_n}) \mid \mathbf{Y}_1,\ldots,\mathbf{Y}_j] = 0 \; a.e., \; \forall j \leq r_n \tag{4}$$

and $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ i.i.d. with a distribution F. Moreover, the $\eta_{n,i_1\cdots i_{m_n}}, 1 \leq i_1 < \cdots < i_{m_n} \leq n$ are such that (2) holds and

$$\sum_{i_1,\ldots,i_{m_n}}^{1,n} \eta^2_{n,i_1\cdots i_{m_n}} = M_n(\nearrow \text{ in } n \geq m_n). \tag{5}$$

Lemma 4.1. *Consider T_n as in (1), with $m_n = r_n + 1$. Define*

$$Z_{nj} = \sum_{i_1,\ldots,i_{m_n-1}}^{1,j-1} \eta_{n,i_1\cdots i_{m_n-1}j}\phi(\mathbf{Y}_{i_1},\cdots,\mathbf{Y}_{i_{m_n-1}},\mathbf{Y}_j),$$

for every $j = m_n,\ldots,n$, and $T_{nk} = Z_{nm_n} + \cdots + Z_{nk}$, for $m_n \leq k \leq n$. Define $\mathcal{B}_{nk} = \sigma\,(\mathbf{Y}_i,\ i \leq k,\ \text{for } m_n \leq k \leq n)$. Then, $\{T_{nk}, \mathcal{B}_{nk} : m_n \leq k \leq n\}$ is a zero-mean martingale array closed on the right by T_n.

Proof This proof is omitted. We refer the reader to the proof of Lemma 1 in [Pinheiro et al. , 2009] for details. □

Let $\tau_{2m_n} = \mathrm{E}\phi^2(\mathbf{Y}_1,\ldots,\mathbf{Y}_{m_n}) > 0$ such that

$$\tau_2 = \lim_{m_n\to\infty} \tau_{2m_n} > 0. \tag{6}$$

In Theorems 4.1 and 4.2 we show that T_n will be asymptotically normal for kernels with finite second moments and a uniform integrability condition. Theorem 4.2 proposes a permutation estimator of τ^2. Theorem 4.3 and 4.4 drop the uniform integrability condition with stronger moment conditions for ϕ, finite $(2+\delta)$-th moments for the former and finite fourth moments for the latter. Both Theorems 4.3 and 4.4 possess the same permutation estimator of τ^2.

The following notation will be used in these theorems. Take the following cumulative sums:

$$m_{nk} = \sum_{i_1,\ldots,i_{m_n-1}}^{1,k} \eta^2_{n,i_1,\ldots,i_{m_n-1},k}, \tag{7}$$

$$\nu_{nk} = m_{nk}\tau_2, \tag{8}$$

$$\nu_n = \nu_{nm_n} + \ldots + \nu_{nn} = M_n\tau_2, \tag{9}$$

when $m_n \leq k \leq n$.

We take

$$\max_{m_n \leq k \leq n} m_{nk}/M_n \to 0 \text{ when } n \to \infty, \tag{10}$$

$$Z^2_{nk}/m_{nk} \text{ are uniformly integrable when } n \to \infty. \tag{11}$$

Theorem 4.1. *Let $\phi(\cdot,\cdot)$ be a degree m_n kernel, centered, stationary of order $m_n - 1$, such that $m_n = o(n)$ as $n \nearrow \infty$, for which* **(A)** *(2), (5) and (6)-(11) hold. Then,*

$$L_n = (\nu_n)^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0,1) \text{ as } n \to \infty. \tag{12}$$

Proof We use Corollary 2.8 from [McLeish , 1974]. Inequality 3.5 from [Burkholder , 1974] is employed to show the first condition of the aforementioned corollary, while $m_n = o(n)$ is needed for its first condition. We refer the reader to Theorem 1 of [Pinheiro et al. , 2011] for further details. □

Theorem 4.2. *Let $\phi(\cdot,\cdot)$ be a degree m_n kernel, centered, stationary of order $m_n - 1$, such that $m_n = o(n)$ as $n \to \infty$, for which* **(A)** *(2), (5) and (6)-(11) hold. Let*

$$U_n^{(m_n)} = \binom{n}{m_n}^{-1} \sum_{i_1,\ldots,i_{m_n}}^{1,n} \phi^2(\mathbf{Y}_{i_1},\ldots,\mathbf{Y}_{i_{m_n}}). \tag{13}$$

Then as $n \to \infty$,

$$L_n = (M_n U_n^{(m_n)})^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0,1). \tag{14}$$

Proof Since $U_n^{(m_n)}$, a U-statistic and unbiased estimator of τ_{2m_n}, is a reverse martingale, $U_n^{(m_n)} \xrightarrow{a.s.} \tau_2$, as $n \to \infty$. (14) then follows from (9) and (12). □

Theorem 4.3. *Let $\phi(\cdot,\cdot)$ be a kernel and assume that, for each n, ϕ has degree m_n (which increases in $n \to \infty$ but $m_n = o(n)$), is centered, and stationary of order $m_n - 1$ such that*

(B.1) $\mathrm{E}\,|\phi(\mathbf{Y}_1,\ldots,\mathbf{Y}_{m_n})|^{2+\delta} < \infty$, *for some positive δ,*

(B.2) *(2), (5) and (6)-(10) hold.*

Let $U_n^{(m_n)}$ *be defined by (13). Then as* $n \to \infty$,

$$L_n = (M_n U_n^{(m_n)})^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0,1). \tag{15}$$

.

Proof Note that $U_n^{(m_n)}$ is an estimator of τ_{2m_n} with bounded variance (away from 0 and ∞). Thus, $U_n^{(m_n)} \xrightarrow{a.s.} \tau_2$, as $n \to \infty$. Since the $(2+\delta)$-th moment of ϕ is finite one does not need the uniform integrability condition, and (15) follows from martingale CLT and Slustky given m_n/n small. □

Now we assume a finite fourth moment of ϕ . This enables us to base our variance estimation on permutation techniques. Take

$$b_n^{(j)} = \sum_{i_1,\ldots,i_{m_n}}^{1,n} \sum_{j_1,\ldots,j_{m_n}}^{1,n} \eta_{n,i_1\ldots i_{m_n}} \eta_{n,j_1\ldots j_{m_n}},$$

for $j = 0, \ldots, m_n$. Note that $\sum_{j=0}^{m_n-1} b_n^{(j)} = -M_n$, and assume

$$\sum_{j=0}^{m_n} b_n^{(j)2} / n^{m_n+j-1} = o(M_n^2) \text{ as } n \to \infty. \tag{16}$$

Theorem 4.4. *Let* $\phi(\cdot,\cdot)$ *be a degree* m_n *kernel, centered, stationary of order* $m_n - 1$ *such that* m_n *increases in* $n \to \infty$ *but* $m_n = o(n)$*, and*

(C.1) $\mathrm{E}\phi^4(\mathbf{Y}_1, \ldots, \mathbf{Y}_{m_n}) < \infty$,

(C.2) *(2), (5), (6)-(9) and (16) hold.*

Then, as $n \to \infty$,

$$L_n = (M_n U_n^{(m_n)})^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0,1). \tag{17}$$

Proof We use the Martingale Array Central Limit Theorem from Dvoretzky [1972]. Instead of the usual Lindeberg condition we employ the (stronger) Liapounouff condition. The degeneracy of cross products of U-statistics and (16) ensure the CLT's conditions. □

In the remaining text we present two theorems for high-dimensional data. For this, besides the increasing degree of ϕ an extra dimensional burden is put on the problem as K may be large. This is not a problem for quasi U-statistics if n is also large, as can be seen in Theorem 4.5, for which no extra conditions are imposed when both n and/or K are large. Theorem 4.6 discusses the case in which K is large and n may be small. Conditions for this latter theorem are quite simplified when compared with the former. In both instances stochastic weights are addressed. Since stochastic weights

may have direct applicability due to sampling schemes this *new* situation is very interesting specially because it doesn't impose any major burden on the proof. Thence, we can consider deterministic weights as a special case.

Theorem 4.5. *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a sequence of i.i.d. $K \times 1$ random vectors. Let $\phi(\cdot, \ldots, \cdot)$ be a kernel of degree m, which increases in $n \to \infty$ but $m = o(n)$, such that*

$$\phi(\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_m}) = \frac{1}{K} \sum_{l=1}^{K} \phi^{\star}(Y_{i_1 l}, \ldots, Y_{i_m l}), \tag{18}$$

for some kernel, stationary of order $m-1$, $\phi^{\star}(\cdot, \ldots, \cdot)$. Let T_n be defined by (1). Assume that one out of the following set of conditions:

(a) *- (6)-(9), (11), $m_{nk} = o_p(M_n)$ as $n \to \infty$ hold ;*

(b) *- (6)-(9), $m_{nk} = o_p(M_n)$ as $n \to \infty$ and* **(B.1)** *hold ;*

(c) *- (6), $\sum_{j=0}^{m} b_n^{(j)2} / n^{m+j-1} = o_p(M_n^2)$ as $n \to \infty$ and* **(C.1)** *hold.*

Suppose that $\{\eta_{n i_1 \ldots i_m},\ 1 \leq i_1 < \cdots < i_m \leq n\ ,\ n \geq m\}$ is a triangular array of random variables independent of $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\ ,\ n \geq m\}$, and

$$\sum_{1 \leq i_1 < \cdots < i_m \leq n} \eta^2_{n i_1 \ldots i_m} - M_n = o_p(M_n) \text{ as } n \to \infty. \tag{19}$$

Suppose also that

$$\sum_{1 \leq l < q \leq K} \mathrm{E}\left[\phi^{\star}(Y_{i_1 l}, \cdots, Y_{i_m l}) \phi^{\star}(Y_{i_1 q}, \cdots, Y_{i_m q})\right] = O(K) \text{ as } K \to \infty. \tag{20}$$

Then

$$(K M_n U_n^{(m)})^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0,1) \text{ as } n \to \infty \text{ and } K \to \infty, \tag{21}$$

where $U_n^{(m)}$ is defined by (13).

Proof Take $\{\eta_{n, i_1 \ldots i_m} : 1 \leq i_1 < \ldots < i_m \leq n\}$, $n \geq m$ to be deterministic, such that $\sum_{i_1, \ldots, i_m}^{1,n} \eta_{n, i_1 \ldots i_m} = 0$ and $\sum_{i_1, \ldots, i_m}^{1,n} \eta^2_{n, i_1 \ldots i_m} = \binom{n}{m}$. Employ Theorems 4.1 and 4.2.

The asymptotic equivalence in probability of stochastic and deterministic weights is then proved given the regularity condition (19), and (21) follows by Dvoretzky [1972]. □

Theorem 4.6. *Let T_n be defined as in Theorem 4.5. Suppose that (20) holds. Then,*

$$T_n / \sqrt{Var(T_n)} \xrightarrow{\mathcal{D}} N(0,1),$$

as $K \to \infty$ (either if $n \to \infty$, $n/K \to 0$, as $K \to \infty$ or if n is bounded).

Proof

We apply Theorem 2.1 [Withers, 1981]. Let

$$S_n = T_n/\sqrt{M_n} = K^{-1}\sum_{k=1}^{K} t_{nk}/\sqrt{M_n} = K^{-1}\sum_{k=1}^{K} x_{nk},$$

where $t_{nk} = \sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\ldots i_m}\phi^*(X_{i_1k},\ldots,X_{i_mk})$.

The rate of growth of the partial sums $(2+\epsilon)$-norm is guaranteed by construction.

The mixing condition (20) ensures the l-mixing [Yoshihara, 1993]. Moreover, (20) also implies that $Var(S_n) = O(K) \to \infty$ as $K \to \infty$ and that the covariances are absolutely summable. Therefore, the CLT holds for T_n at a rate $O(\sqrt{K})$ if n is bounded or $O(n\sqrt{K})$ if both $K \to \infty$ and $n \to \infty$. □

5. Representation of Diversity Measures as Weighted U-statistics and Quasi U-statistics

In this section we propose a class of multidimensional diversity measures. This class has as special cases all the aforementioned within - populations diversity measures. Moreover, we are able to generate both within-populations and between-populations which are functionally equivalent and are asymptotically normal under both null and alternative hypotheses. We motivate this procedure by such a sub-group decomposition of the Hamming distance, as presented in Pinheiro et al. (2005).

Suppose a population that can be naturally divided in G groups. One is interested in assessing information of homogeneity among the groups. Consider n_g observations for group $g = 1,\ldots,G$. Each observation is a K-dimensional vector so that each variate can assume one of C categorical values. For instance, in genetic problems, each dimension can be the response of a particular gene, a DNA basis or some aminoacid, as in proteinomic data. We may assume that K is limited as in Theorem 4.1 or that K is large, such as in Theorems 4.2-4.6.

Let $\mathbf{X}_{g1},\ldots,\mathbf{X}_{gn_g}$ be the K-dimensional i.i.d. observations for the g-th group, $g = 1,\ldots,G$. We can define the gene (Gini-Simpson) diversity measure for the g-th group as

$$\mathcal{D}_g = 1 - \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi^2_{(g)kc}, \tag{22}$$

for $\boldsymbol{\pi}_{(g)} = \{\pi_{(g)kc}\}_{k=1,\ldots,K;c=1,\ldots,C}$, where $\pi_{(g)kc}$ is the probability of $X_{g1k} = c$, $c = 1,\ldots,C$, $k = 1,\ldots,K$, and $g = 1,\ldots,G$.

The Hamming distance between $\mathbf{X}_{gi}$ and $\mathbf{X}_{gj}$ is given by

$$D_{g:ij} = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}(X_{gik} \neq X_{gjk})\ g = 1, \ldots, G.$$

As a non-degenerate U-statistics of degree two, an unbiased and asymptotically normal estimator of $\mathcal{D}_g$ is given by

$$\bar{D}_{gg} = \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} D_{g:ij}\ g = 1, \ldots, G. \tag{23}$$

Similarly to the within-populations measures defined by (23), one naturally defines the between-populations by the generalized two-sample U-statistics

$$\bar{D}_{gg'} = \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} D_{gg':ij}, \tag{24}$$

where

$$D_{gg':ij} = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}(X_{gik} \neq X_{g'jk})\ 1 \leq g < g' \leq G.$$

$\bar{D}_{gg'}$ is an unbiased and asymptotically normal estimator of the Gini-Simpson diversity measure between the g-th and g'-th populations.

$$\mathcal{D}_{gg'} = 1 - \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{(g)kc}\pi_{(g')kc}. \tag{25}$$

Pinheiro et al. (2005) then tests $H_0 : \mathcal{D}_{gg'} = \mathcal{D}_{gg}\ \forall g \neq g'$ versus $H_1 : 2\mathcal{D}_{gg'} > \mathcal{D}_{gg} + \mathcal{D}_{g'g'}$ for some $g < g'$. The following subgroup decomposition of the overall sample Hamming distance id defined:

$$D_n^{(0)} = D_n(W) + D_n(B), \tag{26}$$

where $D_n(W)$ and $D_n(B)$ are the overall within and between measures, respectively.

Under H_0, $\mathrm{E}D_n(B) = 0$ and can be written as a contrast based on a degenerate U-statistics kernel, say ϕ. Eventhough ϕ is degenerate, $(n-1)D_n(B)$ is proven to be asymptotically normal (Pinheiro et al., 2011).

Motivated by this, we propose the following class of diversity class

Definition 5.1 (A Class of Multidimensional Diversity Measures)**.** *The class of U-statistics and Quasi U-statistics estimable and decomposable multidimensional diversity measures, $\mathcal{M}$, is given by any $h(\boldsymbol{\pi}, \mathbf{w})$ such that*

$$h(\boldsymbol{\pi}, \mathbf{w}, \mathbf{w}_*, K) =$$

$$= \mathbf{g}\left(\frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{j=1}^{\infty} w_j \pi_{ck}(1-\pi_{ck})^j, \frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{j=2}^{\infty} w_{j*}\pi_{ck}^j(1-\pi_{ck})\right),$$

where $\mathbf{g}:\mathbb{R}^2\to\mathbb{R}$ *is a continuous and differentiable function,* $\boldsymbol{\pi}$ *is a* $C\times K$ *matrix for which each column is a discrete probability distribution, and* $\mathbf{w}$ *and* $\mathbf{w}_*$ *are (possibly infinite) sequences of constants.*

We should note that we can easily extend this class to contain its limits in the sense that any $h(\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,\infty)$, defined by

$$\mathbf{g}\left(\lim_{K\to\infty}\frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{j=1}^{\infty} w_j \pi_{ck}(1-\pi_{ck})^j, \lim_{K\to\infty}\frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{j=2}^{\infty} w_{j*}\pi_{ck}^j(1-\pi_{ck})\right),$$

should be an element of $\mathcal{M}$ whenever this limit is reasonable. For the sake of notational simplicity, we will continue treating K as finite, unless noted otherwise. Below, we illustrate this class with the generalized classical diversity measures. First, we define

$$\mathcal{H}_{K,j+1} \equiv K^{-1}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}(1-\pi_{ck})^j = \mathrm{E}\left[\phi_{j+1}(\mathbf{X}_1,\ldots,\mathbf{X}_{j+1})\right], \text{ where}$$

$$\phi_{j+1}(\mathbf{X}_1,\ldots,\mathbf{X}_{j+1}) = K^{-1}(j+1)^{-1}\sum_{k=1}^{K}\sum_{i=1}^{j+1}\mathbb{I}\left[X_{ik}\neq X_{lk}, l\neq i\right], \text{ and}$$

$$\mathcal{H}_{K,j+1*} \equiv K^{-1}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}^j(1-\pi_{ck}) = \mathrm{E}\left[\phi_{j+1*}(\mathbf{X}_1,\ldots,\mathbf{X}_{j+1})\right], \text{ where}$$

$$\phi_{j+1*}(\mathbf{X}_1,\ldots,\mathbf{X}_{j+1}) =$$

$$= K^{-1}(j+1)^{-1}\sum_{k=1}^{K}\sum_{i=1}^{j+1}\mathbb{I}\left[X_{i_1k}=\cdots=X_{i_jk}\neq X_{ik}; i_1,\ldots,i_j\neq i\right],$$

for $j=1,\ldots,n-1$.

A suitable combination of $K^{-1}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}(1-\pi_{ck})^j$ and $K^{-1}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}^j(1-\pi_{ck})$, $j\geq 1$ can represent each cited diversity measure, and all of those can be estimated by weighted U-statistics of the form $\binom{n}{j+1}^{-1}\sum_{i_1,\ldots,i_{j+1}}^{1,n}\phi_{j+1}(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_{j+1}})$ and $\binom{n}{j+1}^{-1}\sum_{i_1,\ldots,i_{j+1}}^{1,n}\phi_{j+1*}(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_{j+1}})$, $j\leq 1$. Here $\sum_{i_1,\ldots,i_{j+1}}^{1,n}$ represents the sum for $1\leq i_1<\cdots<i_{j+1}\leq n$.

Consider $r \in \mathbb{N}$. The Shannon index can be written as

$$h_{Sh}(\pi) = h_{Ha}(\pi) + K^{-1} \sum_{j=2}^{\infty} \sum_{k=1}^{K} \frac{1}{j} \pi_{ck}(1-\pi_{ck})^j,$$

which is estimated by the following linear combination of U-statistics as

$$\widehat{h}_{Sh;r} = \sum_{j=1}^{r} j^{-1} \binom{n}{j+1}^{-1} \sum_{i_1,\ldots,i_{j+1}}^{1,n} \phi_{j+1}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{j+1}}).$$

The α-th order entropy of Havrda and Charavát can be written as

$$h_{HC,\alpha} = \left[2^{\alpha-1}-1\right]^{-1} \left[1 - K^{-1} \sum_{k=1}^{K} \sum_{j=1}^{\infty} \binom{\alpha-1}{j} (-1)^j \sum_{c=1}^{C} \pi_{ck}(1-\pi_{ck})^j\right],$$

for which a natural estimator is

$$\widehat{h}_{HC,\alpha;r} =$$

$$= \left[2^{\alpha-1}-1\right]^{-1} \left[1 - \sum_{j=1}^{r} \binom{\alpha-1}{j} (-1)^j \binom{n}{j+1}^{-1} \sum_{i_1,\ldots,i_{j+1}}^{1,n} \phi_{j+1}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{j+1}})\right].$$

The paired Shannon entropy can be represented as

$$h_{pSh}(\pi) =$$

$$= 2h_{Ha}(\pi) + K^{-1} \sum_{j=2}^{\infty} \frac{1}{j} \sum_{k=1}^{K} \sum_{c=1}^{C} \pi_{ck}(1-\pi_{ck})^j + K^{-1} \sum_{j=2}^{\infty} \frac{1}{j} \sum_{k=1}^{K} \sum_{c=1}^{C} \pi_{ck}^j (1-\pi_{ck})$$

so that it is naturally estimated by

$$\widehat{h}_{pSh}(\pi) = 2\binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \phi_2(\mathbf{X}_i, \mathbf{X}_j) +$$

$$+ \sum_{j=2}^{r} j^{-1} \binom{n}{j+1}^{-1} \sum_{i_1,\ldots,i_{j+1}}^{1,n} \phi_{j+1}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{j+1}}) +$$

$$+ \sum_{j=2}^{r} j^{-1} \binom{n}{j+1}^{-1} \sum_{i_1,\ldots,i_{j+1}}^{1,n} \phi_{j+1*}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{j+1}}).$$

In the case of the α-degree entropy of Renyi, note that

$$exp\left(\{(1-\alpha)h_{R,\alpha}(\pi)\}\right) = 1 - (2^{\alpha-1}-1)h_{HC,\alpha}(\pi),$$

so that

$$\widehat{h}_{R,\alpha;r} = 1/(1-\alpha)\ln\left\{1-(2^{\alpha-1}-1)\widehat{h}_{HC,\alpha;r}\right\}.$$

Finally, we can represent the γ-entropy function as

$$h_{E,\gamma}(\pi) = \left(1-2^{\gamma-1}\right)^{-1}\left[1-\left\{1-(2^{1/\gamma-1}-1)h_{HC,\alpha}(\pi)\right\}^{\gamma}\right],$$

and its U-statistics based estimator is given by

$$\widehat{h}_{E,\gamma;r} = 1/\left(1-2^{\gamma-1}\right)\left[1-\left\{1-(2^{1/\gamma-1}-1)\widehat{h}_{HC,\alpha;r}\right\}^{\gamma}\right].$$

In general, $h(\boldsymbol{\pi}, \mathbf{w}, \mathbf{w}_*, K)$ can be estimated by

$$\begin{aligned}\widehat{h}(\mathbf{w}, \mathbf{w}_*, K) \;=\; & \mathbf{g}\left(\sum_{j=1}^{\infty} w_j \binom{n}{j+1}^{-1} \sum_{i_1,\ldots,i_{j+1}}^{1,n} \phi_{j+1}(X_{i_1},\ldots,X_{i_{j+1}}),\right. \\ & \left.\sum_{j=2}^{\infty} w_{j*} \binom{n}{j+1}^{-1} \sum_{i_1,\ldots,i_{j+1}}^{1,n} \phi_{j+1*}(X_{i_1},\ldots,X_{i_{j+1}})\right).\end{aligned}$$

An analogously subgroup decomposability that holds for the generalized Hamming distance given by (26) will hold for any of the aforementioned diversity measures. Therefore, any diversity measure based on Definition 5.1 will provide us with within-populations and between-populations diversity measures (indistinguishably from the functional point of view), a decomposition of the overall measure of dissimilarity, and a unilateral test procedure for homogeneous probabilistic distributions along the groups.

Suppose now G populations driven by G different discrete distributions, say $\boldsymbol{\pi}_{(1)},\ldots,\boldsymbol{\pi}_{(G)}$, and from which a sample of n observations is drawn as follows. $\mathbf{X}_{g,1},\ldots,\mathbf{X}_{g,n_g}$ are i.i.d. $\sim \boldsymbol{\pi}_{(g)}$, $g=1,\ldots,G$, $n = n_1+\cdots+n_G$. One can build a between-populations measure of diversity which will be a contrast of U-statistics, based upon functions of the previously mentioned kernels ϕ_{j+1}, $j=1,2,\ldots$ and ϕ_{l+1*}, $l=2,\ldots$, say $\phi^{(0)}_{j+1}$, $j=1,2,\ldots$, and $\phi^{(0)}_{l+1*}$, $l=2,\ldots$. Even though, under the hypothesis of homogeneity, $H_0 : \boldsymbol{\pi}_{(1)} = \cdots = \boldsymbol{\pi}_{(G)}$, the aforementioned kernels $\{\phi^{(0)}$ and $\phi_{l*}(0)$ are all degenerate (of stationary order 1), we are able to prove asymptotic normality, by the general results in Section 2.

We will prove now that such estimators for the within-populations diversity measures will be asymptotically normal whenever at least one of the kernels is non-degenerate. The non-degeneracy is true unless the discrete probability distribution is uniform for every $k=1,\ldots,K$ or if every characteristic equals one category with probability one (Pinheiro et al., 2008).

Other than those rather uninteresting exceptions, the usual U-statistics asymptotics will prevail for n large and limited K. In genetic analysis, one usually has large K (Sen, 2006). If n and K are both large, asymptotic normality will be also true for any dependent structure considered along the K characteristics. If n is limited but K large, it is proven that normal asymptotics will hold, under reasonable mixing conditions. However, for large K and small n, if the degree of the kernel is infinite, a bias will remain, giving the projection of the kernel onto a maximum degree (limited by n).

We first consider the within-populations measures. For that, we need to employ Hoeffding's decomposition on the U-statistics. However, since, in the case of homogeneity tests, we will be dealing with first-order degeneracy, we present the following set of orthonormal functions for both Hoeffding's projection on both their first and second orders.

Without loss of generality, we will present the following results with no indexing by groups. Whenever those differences matter, we will point that out and proceed accordingly. The first and secod order Hoeffding projections are given by:

$$\psi_{21}(\mathbf{x}_1) = \mathrm{E}\left(\phi_2(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1 = \mathbf{x}_1\right) \;=\; K^{-1}\sum_{k=1}^{K}(1 - \pi_{x_{1k}k}),$$

$$\psi_{22}(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{E}\left(\phi_2(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2\right) \;=\; K^{-1}\sum_{k=1}^{K}\mathbb{I}(x_{1k} \neq x_{2k}),$$

$$\psi_{j+1,1}(\mathbf{x}_1) = (j{+}1)^{-1}K^{-1}\left\{\sum_{k=1}^{K}(1 - \pi_{x_{1k}k})^j + j\sum_{k=1}^{K}\sum_{d\neq x_{1k}}\pi_{dk}(1 - \pi_{dk})^{j-1}\right\}$$
$$\text{for } 2 \leq j \leq r,$$

$$\psi_{j+1,2}(\mathbf{x}_1, \mathbf{x}_2) = (j+1)^{-1}K^{-1}\left\{\sum_{k=1}^{K}(1 - \pi_{x_{1k}k})^{j-1}\mathbb{I}(x_{1k} \neq x_{2k}) + \right.$$
$$\left. + \sum_{k=1}^{K}(1 - \pi_{x_{2k}k})^{j-1}\mathbb{I}(x_{1k} \neq x_{2k}) + (j-1)\sum_{k=1}^{K}\sum_{c\neq x_{1k},x_{2k}}\pi_{ck}\left(1 - \pi_{ck}\right)^{j-2}\right\}$$
$$\text{for } 2 \leq j \leq r,$$

$$\psi_{j+1,1*}(\mathbf{x}_1, \mathbf{x}_2) = (j+1)^{-1}K^{-1}\left\{j\sum_{k=1}^{K}\pi_{x_{1k}k}^{j-1}(1 - \pi_{x_{1k}k}) + \sum_{k=1}^{K}(1 - \pi_{x_{1k}k}^{j})\right\},$$

$$\psi_{j+1,2*}(\mathbf{x}_1,\mathbf{x}_2)=(j+1)^{-1}K^{-1}\left\{\sum_{k=1}^{K}\pi_{x_{1k}k}^{j-1}\mathbb{I}(x_{1k}\neq x_{2k})+\right.$$

$$\left.+\sum_{k=1}^{K}\pi_{x_{2k}k}^{j-1}\mathbb{I}(x_{1k}\neq x_{2k})+(j-1)\sum_{k=1}^{K}\pi_{x_{1k}k}^{j-2}(1-\pi_{x_{1k}k})\mathbb{I}(x_{1k}\neq x_{2k})\right\}$$

$$\text{for } 2\leq j\leq r.$$

Let $\pi_{ck,dl}$ be the probability of observing a c on the k-th position, and d on the l-th position. Since we do not impose any dependency structure these and other probabilities need to be taken into account as such. Consider

$$h(\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K)=\mathbf{g}\left(\theta(\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K)\right),$$

where

$$\boldsymbol{\theta}(\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K)=$$

$$=\left(\frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{j=1}^{\infty}w_j\pi_{ck}(1-\pi_{ck})^j,\frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{j=2}^{\infty}w_{j*}\pi_{ck}^{j}(1-\pi_{ck})\right).$$

We have

$$\mathrm{E}\psi_{21}(\mathbf{X}_i)\psi_{j+1,1}(\mathbf{X}_i)=K^{-2}(j+1)^{-1}\left\{\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}(1-\pi_{ck})^j\left[1-(j+1)\pi_{ck}\right]\right.$$

$$+j\sum_{k=1}^{K}\sum_{c,d}\pi_{ck}\pi_{dk}(1-\pi_{ck})(1-\pi_{dk})^{j-1}+\sum_{k\neq l}\sum_{c,d}\pi_{ck,dl}(1-\pi_{ck})(1-\pi_{dl})^{j-1}$$

$$\left.\times\left[1-(j+1)\pi_{dl}\right]+j\left[K^2\mathcal{H}_{K,2}\mathcal{H}_{K,j-1}-\sum_{k=1}^{K}\sum_{c,d}\pi_{ck}\pi_{dk}(1-\pi_{ck})(1-\pi_{dl})^{j-1}\right]\right\}.$$

$$\mathrm{E}\psi_{j+1,1}(\mathbf{X}_i)^2=(j+1)^{-2}K^{-2}\left\{\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}(1-\pi_{ck})^{2(j-1)}\times\right.$$

$$\left[(1-\pi_{ck})^2-2j\pi_{ck}(1-\pi_{ck})+j^2\pi_{ck}^2\right]+j\sum_{k=1}^{k}\sum_{d=1}^{C}\pi_{dk}(1-\pi_{dk})^{j-1}\sum_{c=1}^{C}\pi_{ck}\times$$

$$\left[2(1-\pi_{ck})^j+j\sum_{d=1}^{C}\pi_{dk}(1-\pi_{dk})^{j-1}-2j\pi_{ck}(1-\pi_{ck})^{j-1}\right]+$$

$$+\sum_{k\neq l}\sum_{c,d}\pi_{ck,dl}(1-\pi_{ck})^j(1-\pi_{dl})^j+\sum_{k\neq l}\sum_{c,d,e}\pi_{ck,dl}(1-\pi_{ck})^j\pi_{el}(1-\pi_{el})^{j-1}-$$

$$-2j\sum_{k\neq l}\sum_{c,d}\pi_{ck,dl}(1-\pi_{ck})^{j}\pi_{dl}(1-\pi_{dl})^{j-1}+$$

$$+j^2\sum_{k\neq l}\sum_{c,d,e,f}\pi_{ck,dl}\pi_{el}\pi_{fk}(1-\pi_{el})^{j-1}(1-\pi_{fk})^{j-1}-$$

$$-j^2\sum_{k=1}^{K}\sum_{c,d,f}\pi_{ck,dl}\pi_{dl}\pi_{fk}(1-\pi_{dl})^{j-1}\times$$

$$\left.\times(1-\pi_{fk})^{j-1}-j^2\sum_{k=1}^{K}\sum_{c,d}\pi_{ck,dl}\pi_{ck}\pi_{dl}(1-\pi_{ck})^{j-1}(1-\pi_{dl})^{j-1}\right\}.$$

$$\mathrm{E}\psi_{j+1,1}(\mathbf{X}_i)\psi_{m+1,1}(\mathbf{X}_i)=K^{-2}(j+1)^{-1}(m+1)^{-1}\times$$

$$\left\{\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}(1-\pi_{ck})^{j}\sum_{d\neq c}\pi_{dk}(1-\pi_{dk})^{m-1}\right.$$

$$+m\sum_{k\neq l}\sum_{c,d}\pi_{ck,dl}(1-\pi_{ck})^{j}\sum_{e\neq d}\pi_{el}(1-\pi_{el})^{m-1}+j\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}(1-\pi_{ck})^{m-1}$$

$$\times\sum_{d\neq c}\pi_{dk}(1-\pi_{dk})^{j-1}+j\sum_{k\neq l}\sum_{c,d}\pi_{ck}(1-\pi_{ck})^{m}\sum_{e\neq d}\pi_{el}(1-\pi_{el})^{j-1}$$

$$+jm\sum_{k\neq l}\sum_{c=1}^{C}\pi_{ck}\sum_{d\neq c}\pi_{dk}(1-\pi_{dk})^{j-1}\sum_{e\neq c}\pi_{ek}(1-\pi_{ek})^{m-1}$$

$$\left.+jm\sum_{k\neq l}\sum_{c,d}\pi_{ck,dl}\sum_{e\neq c}\pi_{ek}(1-\pi_{ek})^{j-1}\sum_{f\neq d}\pi_{fl}(1-\pi_{fl})^{m-1}\right\}.$$

$$\mathrm{E}\psi_{j+1,1}(\mathbf{X}_i)\psi_{m+1,1*}(\mathbf{X}_i)=K^{-2}(j+1)^{-1}(m+1)^{-1}\times$$

$$\times\left\{\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}(1-\pi_{ck})^{j}\left[(1-\pi_{ck})^{m}\,mj\pi_{ck}^{m}-j\pi_{ck}(1-\pi_{ck})^{m-1}+\right.\right.$$

$$\left.+m\pi_{ck}^{m-1}(1-\pi_{ck})\right]+j\sum_{k=1}^{K}\sum_{c,d=1}^{C}\pi_{ck}\pi_{dk}(1-\pi_{ck})^{m}(1-\pi_{dk})^{j-1}$$

$$+mj\sum_{k=1}^{K}\sum_{c,d=1}^{C}\pi_{ck}^{m}\pi_{dk}(1-\pi_{ck})(1-\pi_{dk})^{j-1}+\sum_{k\neq l}\sum_{c,d=1}^{C}\pi_{ck,dl}(1-\pi_{ck})^{j}(1-\pi_{dl})^{m}+$$

$$+m\sum_{k\neq l}\sum_{c,d=1}^{C}\pi_{ck,dl}(1-\pi_{ck})^{j}\pi_{dl}^{m-1}(1-\pi_{dl})+$$

$$+j\sum_{k\neq l}\sum_{c,d,e=1}^{C}\pi_{ck,dl}\pi_{ek}(1-\pi_{ek})^{j-1}(1-\pi_{dl})^{m}$$

$$+mj\sum_{k\neq l}\sum_{c,d,e=1}^{C}\pi_{ck,dl}\pi_{ek}(1-\pi_{ek})^{j-1}\pi_{dl}^{m-1}(1-\pi_{dl})-$$

$$-j\sum_{k\neq l}\sum_{c,d=1}^{C}\pi_{ck,dl}\pi_{ck}(1-\pi_{ck})^{j-1}(1-\pi_{dl})-$$

$$-mj\sum_{k\neq l}\sum_{c,d=1}^{C}\pi_{ck,dl}\pi_{ck}(1-\pi_{ck})^{j-1}\pi_{dl}^{m-1}(1-\pi_{dl})^{m}\Bigg\}.$$

$$\mathrm{E}\psi_{j+1,1*}(\mathbf{X}_i)^2=(j+1)^{-2}K^{-2}\Bigg\{j^2\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}^{2j-1}(1-\pi_{ck})^2+$$

$$+j^2\sum_{k\neq l}\sum_{c,d=1}^{C}\pi_{ck,dl}\pi_{ck}^{j-1}\pi_{dl}^{j-1}(1-\pi_{ck})(1-\pi_{dl})+2j\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}^{j}(1-\pi_{ck})^{j+1}+$$

$$+2\frac{j}{(j+1)^2}K^{-2}\sum_{k\neq l}\sum_{c,d=1}^{C}\pi_{ck,dl}\pi_{ck}^{j-1}(1-\pi_{ck})(1-\pi_{dl})^{j}+$$

$$+\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}(1-\pi_{ck})^{2j}+\sum_{k\neq l}\sum_{c,d=1}^{C}\pi_{ck,dl}(1-\pi_{ck})^2(1-\pi_{dl})^{j}\Bigg\}.$$

$$\mathrm{E}\psi_{21}(\mathbf{X}_i)\psi_{m+1,1*}(\mathbf{X}_i)=(m+1)^{-2}K^{-2}\Bigg\{m\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{ck}^{m}(1-\pi_{ck})^2+$$

$$+\sum_{k\neq l}\sum_{c=1}^{C}\pi_{ck}(1-\pi_{ck})^{m+1}+m\sum_{k\neq l}\sum_{c,d=1}^{C}\pi_{ck,dl}(1-\pi_{ck})\pi_{dl}^{m-1}(1-\pi_{dl})+$$

$$+\sum_{k\neq l}\sum_{c,d=1}^{C}\pi_{ck,dl}(1-\pi_{ck})(1-\pi_{dl})\Bigg\}.$$

Finally,

$$\begin{aligned}
\sigma_{j+1}^2 &= \mathrm{E}\psi_{j+1,1}(\mathbf{X}_1)^2 - \mathcal{H}_{K,j+1}^2, \\
\sigma_{j+1,m+1} &= \mathrm{E}\psi_{j+1,1}(\mathbf{X}_1)\psi_{m+1,1}(\mathbf{X}_1) - \mathcal{H}_{K,j+1}\mathcal{H}_{K,m+1}, \\
\sigma_{j+1*}^2 &= \mathrm{E}\psi_{j+1,1*}(\mathbf{X}_1)^2 - \mathcal{H}_{K,j+1*}^2, \\
\sigma_{j+1,m+1*} &= \mathrm{E}\psi_{j+1,1}(\mathbf{X}_1)\psi_{j+1,1*}(\mathbf{X}_1) - \mathcal{H}_{K,j+1}\mathcal{H}_{K,m+1*}.
\end{aligned}$$

We can then write

$$\begin{aligned}
&Var\left(\sum_{j=1}^{r} w_j \binom{n}{j+1}^{-1} \sum_{i_1,\ldots,i_{j+1}}^{1,n} \phi_{j+1}(X_{i_1},\ldots,X_{i_{j+1}})\right. \\
&\left. + \sum_{j=2}^{\infty} w_j \binom{n}{j+1}^{-1} \sum_{i_1,\ldots,i_{j+1}}^{1,n} \phi_{j+1*}(X_{i_1},\ldots,X_{i_{j+1}}\right) \\
&= \frac{4}{n}\left[\sum_{j=1}^{r} \frac{w_j^2}{(j+1)^2}\sigma_{j+1}^2 + \sum_{j=2}^{r} \frac{w_{j*}^2}{(j+1)^2}\sigma_{j+1*}^2\right. \\
&\left. + 2\sum_{1\le j<m\le r} \frac{w_j w_{m*}}{(j+1)(m+1)}\sigma_{j+1,m+1*}\right] + O(n^{-2}). \qquad (27)
\end{aligned}$$

Note that the sums in the RHS of (27) converge as $r \to \infty$. Let

$$\mathbf{Y}_{ni} = 2\left(\sum_{j=2}^{r} j^{-1} w_j \left[\psi_{j,1}(X_i) - \mathcal{H}_{K,j}\right], \sum_{j=2}^{r} j^{-1} w_{j*}\left[\psi_{j*,1}(X_i) - \mathcal{H}_{K,j*}\right]\right).$$

It is easy to show that

$$\left(\sum_{i=1}^{n} \mathrm{E}|Y_{k,ni} - \mathrm{E}Y_{l,ni}|^3\right)^2 = O(n^2) = o(n^3) = o\left(\sum_{i=1}^{n} Var(Y_{l,ni})\right)^3,$$

as $n \to \infty$, for $l = 1, 2$ (and $r \to \infty$). Therefore,

$$n^{-1/2}\sum_{i=1}^{n}(\mathbf{Y}_{ni} - \mathrm{E}\mathbf{Y}_{ni}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \Sigma_{\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K}),$$

as $n \to \infty$, $r\ (\le n-1) \to \infty$, and the asymptotic variance of $\widehat{\boldsymbol{\theta}}(\boldsymbol{\pi}, \mathbf{w}, \mathbf{w}_*)$ is given by

$$\Sigma_{\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K} =$$

$$= 4 \begin{bmatrix} \sum_{j=1}^{\infty} \frac{w_j^2}{(j+1)^2} \sigma_{j+1}^2 & \sum_{2 \le j < m \le \infty} \frac{w_j w_m}{(j+1)(m+1)} \sigma_{j+1,m+1*} \\ \sum_{2 \le j < m \le \infty} \frac{w_j w_m}{(j+1)(m+1)} \sigma_{j+1,m+1*} & \sum_{j=2}^{\infty} \frac{w_j^2}{(j+1)^2} \sigma_{j+1*}^2 \end{bmatrix}.$$

Moreover, each component of $\Sigma_{\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K}$ converges as $K \to \infty$. Theorem 5.1 summarizes the results for large n.

Theorem 5.1. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random vectors distributed according to the probability distribution $\boldsymbol{\pi}$. Suppose that $\boldsymbol{\pi}$ is such that $\phi_j(\cdot)$ and $\phi_{j,*}(\cdot)$, $j = 1, \ldots$ are all nondegenerate kernels. Then,*

$$\sqrt{n}\left(\widehat{h}(\mathbf{w}, \mathbf{w}_*, K) - h(\boldsymbol{\pi})(\boldsymbol{\pi}, \mathbf{w}, \mathbf{w}_*, K)\right) \xrightarrow{\mathcal{D}} N(0, \gamma^2_{\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K}),$$

as n(and $r \le n-1$) $\to \infty$ but K is limited, where $\dot{g}$ is the gradient of g , and

$$\gamma^2 = \dot{g}(h_{\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K})' \Sigma_{\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K} \dot{g}(h_{\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K}).$$

If n(and $r \le n-1$) $\to \infty$ and $K \to \infty$ such that $Var(\widehat{h}(\mathbf{w}, \mathbf{w}_*, K))$ is of order $O(nv(K)$ as $n, K \to \infty$, then

$$\sqrt{v(K)n}\left(\widehat{h}(\mathbf{w}, \mathbf{w}_*, K) - h(\boldsymbol{\pi}, \mathbf{w}, \mathbf{w}_*, K)\right) \xrightarrow{\mathcal{D}} N(0, \Sigma_{\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,\infty}),$$

as n, K(and $r \le n-1$) $\to \infty$, where

$$\Sigma_{(\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,\infty)} = \lim_{n,K \to \infty} v(K) \Sigma_{\boldsymbol{\pi},\mathbf{w},\mathbf{w}_*,K}.$$

In the case the sample size is not large but the number of sites (or characteristics) can be taken as large, under suitable mixing conditions, one can a similar result but there is a drawback. It forks fine for fixed degree kernels (Pinheiro et al., 2011), but for symmetric kernels a bias term which decreases with n will remain.

6. Conclusion

We propose a unified approach to a large class of diversity measures in the literature. The approach is fully based on symmetric kernels and U-statistics related procedures. Given the characteristics of these diversity measures, one is able to decompose each one in within and between-population dissimilarity measures, naturally leading to a homogeneity test. Moreover, asymptotic normality is proven, under very mild conditions, for both within and between-population measures.

References

Anselmo, C.A.F. and A. Pinheiro (2012). Philogenetic Trees via Hamming Distance Decomposition Tests. *Journal of Statistical Computation and Simulation*, 89, 1287–1297.

Bowman. E.O., E. Hutcheson, E.P. Odum and L.R. Shenton, (1971). Comments on the distribution of indices of diversity. In: *Statistical Ecology* **3** (Patil, G.P., E.C. Pielou and W.E. Waters, eds.).

Burkholder, D.L., Distribution function inequalities for martingales. *Ann. Prob.*, 1(1), 19–42.

Dress, A. and M. Steel (2007). Phylogenetic diversity over an Abelian group. *Annals of Combinatorics* **11**, 143–160.

Dvoretzky, A., Central limit theorem for dependent random variables, Proceedings Sixth Berkeley Symposium Math. Statist. Prob. (Ed. L.LeCam et al.) Los Angeles: University of California Press, Vol.2, 513–555 (1972).

Frees, E.W. (1989). Infinite order U-statistics. *Scandinavian Journal of Statistics* **16**(1), 29–45.

Gilbert, P.B.; A.J. Rossini and R. Shankarappa (2005). Two-sample tests for comparing intra-individual genetic sequence diversity between populations. *Biometrics* **61**, 106–117.

Gillet, E.M. (2007). Qualified estimation of two measures of diversity within populations. Biometrical Journal **49**(2), 272–285.

Gillooly, J.F.; A.P. Allen; G.B. West and J.H. Brown (2005). The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proceedings of the Natinal Academy of Sciences* **102**(1), 140–145.

Gini, C. W. (1912). Variabilita e mutabilita. *Studi Economico-Giuridici della R. Universita di Cagliari* **3**(2), 3-159.

Hill, M. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology* **54**, 427-431.

Hutcheson, K. (1970). A test for comparing diversities based on the Shannon formula. *Journal of Theoretical Biology* **29**, 151-154.

Kim, J-K., Y-S. Jung, E.A.Sungur, K-H. Han, C. Park and I. Sohn (2008). A copula method for modeling directional dependence of genes. *BMC Bioinformatics* **9**:225.

Kussell, E. and S. Leibler (2005). Phenotypic diversity, population growth, and information in fluctuating environments. *Science* **309**, 2075–2078.

Lewontin, R. (1972). The apportionment of human diversity. *Evolutionary Biology* **6**, 391-398.

Magurran, A.E. (1988). *Ecological Diversity and Its Measurement*. Princeton University Press.

Mahalanobis (1936). On the generalized distance in statistics. Proceedings of the National Institute of Science of India **2**, 49–55.

McLeish, D.L. (1974). Dependent central limit theorems and invariance principles. *Ann. Prob.*, 2(4), 620–628.

Moulton, V.; C. Stemple and M. Steel (2007). Optimizing phylogenetic diversity under constraints. *Journal of Theoretical Biology* **246**, 186–194.

Nayak, T.K. and J.L. Gastwirth (1989). The use of diversity analysis to assess the relative influence factors affecting the income distribution. *Journal of Business and Economic Statistics* **7**(4), 453–460.

Nei, M. (1972). Genetic distance between populations, *American Naturalist* **106**, 283-292.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583-590.

Nei, M. and A. Roychoudhury (1974). Sampling variance of heterozygosity and genetic distance. *Genetics* **76**, 379-390.

Peet, R.K. (1974). The measurement of species diversity, *Annual Review of Ecology, Evolution and Systematics* **5**, 285-307.

Pinheiro, H.P., A. Pinheiro and P.K. Sen (2005). Comparison of genomic sequences using the Hamming distance. *Journal of Statistical Planning and Inference* **130**(1-2), 325-339.

Pinheiro, A., Sen, P.K. and Pinheiro, H.P. (2006). Parametric modelling of genomic sequences distance. *Calcutta Statistical Association Bulletin* **58**(229-230) 1-14.

Pinheiro, A., H.P. Pinheiro and S. Kiihl (2008). An asymptotically normal test for the selective neutrality hypothesis. *IMS Collections - Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* **1**, 377-389.

Pinheiro, A., P.K. Sen and Pinheiro, H.P. (2009). Decomposability of high-dimensional diversity measures: quasi U-statistics, martingales and non-standard asymptotics. *Journal of Multivariate Analysis*, **100**, 1645–1656.

Pinheiro, A., P.K. Sen and Pinheiro, H.P. (2011). A class of asymptotically normal degenerate quasi U-statistics. *Annals of the Institute of Statistical Mathematics*, **63**(6), 1165-1182.

Politi, A., M. Moné, A . Houtsmuller , D . Hoogstraten , W . Vermeulen , R . Heinrich , R . van Driel (2005). Mathematical modeling of nucleotide excision repair reveals efficiency of sequential assembly strategies. *Molecular Cell* **19**(5), 679–690.

Rao, C.R. (1982a). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology* **21**, 24-43.

Rao, C.R. (1982b). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankya A* **44**, 1-21.

Rao, C.R. (1982c). Gini-Simpson index of diversity: A characterization, generalization and applications. *Utilitas Mathematica* **21**, 273-282.

Rao, C.R. and R. Boudreau (1984). Diversity and cluster analysis of blood group data on some human populations. In: *Multivariate Statistical Methods in Physical Anthropology* (Van Vark, G and W. Howell, eds.). Reidel, Dordrecht.

Salicrú, M.; S. Vives and J. Ocaña (2005). Testing the homogeneity of diversity measures: a general framework. *Journal of Statistical Planning and Inference* **132**(1-2), 117–129.

Sen, P.K. (1999). Utility-oriented Simpson-type indexes and inequality measures. *Calcutta Statistical Association Bulletin* **49**, 1-22.

Sen, P.K. (2006). Robust statistical inference for high-dimensional data models with application to genomics. *Austrian Journal of Statistics* **35**(2-3), 197–214.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**, 37917423, 62317656.

Simpson, E. H. (1949). The measurement of diversity. *Nature* **163**, 688.

Tavaré, S. and BW Giddings (1989). Some statistical aspects of the primary structure of nucleotide sequences. In: *Mathematical Methods for DNA Sequences* (Waterman, M., ed), 116–132. CRC Press, Boca Raton, FL.

Valk, M. and A. Pinheiro (2012). Time Series Clustering Via Quasi U-Statistics. *Journal of Time Series Analysis*, **33**, 608-619.

Withers, C.S. (1981). Central Limit Theorems for Dependent Variables. I. *Z. Wahrsch. und Verw. Gebiete* **57**(4) 509–534.

Williams, C.B. (1945). Index of diversity as applied to ecological problems. *Nature* **155**, 390-391.

Yoshihara, K. (1993). Asymptotic statistics based on weakly dependent data. *Weakly Dependent Stochastics Sequences and Their Applications, 2.* Sanseido, Tokyo.