Summer Statistics Research
Week 3 Update
Supervisor: Julia Palacios
Co-supervisor: Noah Rosenberg

Alan Aw
alanaw1@stanford.edu
Date: July 12, 2017

Introduction

In population genetics, standard models provide null models against which statistical tests for neutrality can be performed. Under the infinite sites Wright-Fisher model, the scaled mutation rate $\theta = 2pN_e\mu$ — where μ is the mutation rate for the locus, N_e is the effective population size and p is the ploidy number of the cell — forms the basis of statistical tests for neutrality in a locus undergoing evolution. These statistical tests rely on the frequency spectrum of a collection of genetic samples, which is an important summary statistic that can be used to infer mutation rates and population size dynamics in parametric inference, or to test for neutrality in the case of categorical inference. While classical tests such as Tajima's D, Fu and Li's F and Fay and Wu's H have been used widely in testing alternative hypotheses regarding the evolutionary history of genetic data, much discussion continues to surround the inaccuracies of such tests based on genomic datasets arising from next generation sequencing (NGS). Errors from ascertainment biases in SNP calling [3], low sequencing depth [10] and sequence pooling [6] can result in underestimation of nucleotide polymorphisms that appear in very low frequencies in a sample. On the other hand, the statistical power of classical tests turns out not to be optimal in all scenarios: a simulation study and theoretical approach [5] showed that each of the classical tests has a specific frequency spectrum for which it is most powerful, and for other frequency spectra there are optimal difference statistic tests that maximize the statistical power.

Recent studies on the mathematical properties of population genetic statistics have shown that many summary statistics — including F_{ST} [2, 9], homozygosity [12, 13] and its derived statistics [8] — are constrained both by closely related statistics and parameters of the data, so that values of the latter quantities restrict the range of values that the summary statistic can take. These results provide guidance to the usage and interpretation of summary statistics in population genetic studies, by clarifying observations about the behavior of summary statistics with regard to population parameters (e.g., [2]) or with respect to other statistics (e.g., [13]). Moreover, at times they lead to novel empirical insights (e.g., [8]).

In this paper, we interrogate the usefulness of studying mathematical properties of neutrality test statistics and estimators of mutation rates in the infinite sites model. We illustrate how mathematical bounds between neutrality test statistics and components of the frequency spectrum can constrain the range of values for which the test statistic can take. We apply these mathematical bounds to both real data as well as frequency spectra simulated from evolutionary models. Our study not only contributes to the development of a robust statistical framework for neutrality tests based on the frequency spectrum for NGS data, but also advances mathematical approaches to the study of population genetic statistics.

Theory

Throughout this paper, we let $a_n = \sum_{i=1}^{n-1} 1/i$ denote the $(n-1)$th partial sum of the harmonic series and $A_n = \sum_{i=1}^{n-1} 1/i^2$ denote the sum of squares of the first $n-1$ positive integers. Let $\Delta_n := \{\mathbf{x} \in \mathbb{R}^n : x_i \geqslant 0, \sum_i x_i = 1\}$ denote the $(n-1)$-dimensional probability simplex. In a population of p-ploids evolving neutrally according to the infinite sites model, the population-scaled mutation rate (sometimes referred to as nucleotide diversity) $\theta = 2pN_e\mu$ admits a family of unbiased estimators based on the frequency spectrum, which is given by

$$\hat{\theta}_\omega = \sum_{i=1}^{n-1} \omega_i i \xi_i, \tag{1}$$

where $(\omega_1, \ldots, \omega_{n-1}) \in \Delta_{n-1}$. Each estimator of the form given in eq. (1) is indeed unbiased, since under the infinite sites model $E[\xi_i] = \theta/i$ and $\hat{\theta}_\omega$ is a linear combination of weights summing to 1. Note that

$\sum_{i=1}^{n-1} \xi_i = S_n$, the number of segregating sites corresponding to the frequency spectrum.

Given any two choice of weights $\omega_1 = (\omega_{11}, \ldots, \omega_{1(n-1)})$ and $\omega_2 = (\omega_{21}, \ldots, \omega_{2(n-1)})$, we can define the normalized difference statistic T_Ω as the difference between two unbiased θ-estimators $\hat{\theta}_{\omega_1}$ and $\hat{\theta}_{\omega_2}$.

$$T_\Omega = \frac{\hat{\theta}_{\omega_1} - \hat{\theta}_{\omega_2}}{\sqrt{\mathrm{Var}\left(\hat{\theta}_{\omega_1} - \hat{\theta}_{\omega_2}\right)}} = \frac{\sum_{i=1}^{n-1} \Omega_i i \xi_i}{\sqrt{\alpha_n \theta + \beta_n \theta^2}}, \tag{2}$$

where $\Omega_i \stackrel{\text{def}}{=} \omega_{1i} - \omega_{2i}$ and

$$\begin{aligned} \alpha_n &= \sum_{i=1}^{n-1} i\Omega_i^2 \\ \beta_n &= \sum_{i=1}^{n-1} i^2 \Omega_i^2 \sigma_{ii} + 2 \sum_i \sum_{j>i} ij \Omega_i \Omega_j \sigma_{ij} \end{aligned}$$

with σ_{ii} and σ_{ij} — assuming $i > j$ — defined by the following rule:

$$\sigma_{ii} = \begin{cases} b_n(i+1) & \text{if } i < n/2 \\ \frac{2(a_n - a_i)}{n-i} - \frac{1}{i^2} & \text{if } i = n/2 \\ b_n(i) - \frac{1}{i^2} & \text{if } i > n/2 \end{cases}$$

and

$$\sigma_{ij} = \begin{cases} \frac{b_n(i+1) - b_n(i)}{2} & \text{if } i + j < n \\ \frac{a_n - a_i}{n-i} + \frac{a_n - a_j}{n-j} - \frac{b_n(i) + b_n(j+1)}{2} - \frac{1}{ij} & \text{if } i + j = n \\ \frac{b_n(j) - b_n(j+1)}{2} - \frac{1}{ij} & \text{if } i + j > n \end{cases}$$

Note that $b_n(i) \stackrel{\text{def}}{=} n(a_{n+1} - a_i) / \binom{n-i+1}{2} - 2/(n-i)$. These formulas are given in [1, pg. 251]. This framework turns out to generalize the family of classical tests for neutrality, including Tajima's D, Fay and Wu's H and Fu and Li's F. Ferretti et al. (see [5]) demonstrated how this framework allows one to design optimal tests for neutrality, in the sense that the test is most powerful against a known alternative evolutionary model. They also used simulations to obtain frequency spectra for which the classical tests yield maximum power.

In calculating the test statistic T_Ω, one has to estimate θ and θ^2 in the variance expression appearing in the denominator (see RHS of eq. (2)). Because the Watterson estimator θ_W is consistent — $\mathrm{Var}(\theta_W) \to 0$ as the number of samples $n \uparrow \infty$ (see [4, pg. 35]) — it is typically used in this calculation. It can be verified that $E[S_n] = a_n \theta$ and $E[S_n^2 - S_n] = (a_n^2 + A_n)\theta^2$ (see eq. (2.30) in [4, pg. 66]), hence in practice one would obtain T_Ω by setting the denominator $\sqrt{\mathrm{Var}\left(\hat{\theta}_{\omega_1} - \hat{\theta}_{\omega_2}\right)} = \sqrt{\alpha_n \theta + \beta_n \theta^2}$ of eq. (2) to $\sqrt{\alpha_n S_n / a_n + \beta_n S_n (S_n - 1)/(a_n^2 + A_n)}$. Another method of obtaining the denominator is to simultaneously perform simulations on the standard model with $\hat{\theta}_{\omega_1}$ and $\hat{\theta}_{\omega_2}$ as the parameters, and then computing the observed variance.

Results

In this section, we prove our main theorems, which can be viewed as a family of results demonstrating how values of estimators $\hat{\theta}_\omega$ and of the test statistic T_Ω are constrained by any single component ξ_i of the frequency spectrum. Throughout, we assume that we are given n data samples that give rise to the unfolded frequency spectrum $\boldsymbol{\xi} = (\xi_i)_{i=1}^{n-1}$, and moreover the observed number of segregating sites $S_n = \sum_{i=1}^{n-1} \xi_i$. Our proofs rely on the rearrangement inequality of Hardy, Littlewood and Pólya, which is reproduced below. We make use of the notation $\mathcal{S}_n$ to denote the symmetric group acting on a set of n objects.

Proposition 1 (Rearrangement Inequality ([11], pg. 207)). *For any two increasing sequences $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ of real numbers and any permutation $\sigma \in \mathcal{S}_n$,*

$$\sum_{i=1}^{n} x_i y_i \geqslant \sum_{i=1}^{n} x_{\sigma(i)} y_i \geqslant \sum_{i=1}^{n} x_i y_{n-i+1}.$$

Moreover, if $x_1 < \cdots < x_n$ and $y_1 < \cdots < y_n$, then the upper bound is uniquely attained for the identity permutation $\sigma(i) = i$ while the lower bound is uniquely attained for the permutation reversing the order, i.e., $\sigma(i) = n - i + 1$.

Note that Proposition 1 can be viewed as follows: given any two arbitrary real vectors of the same length, the way to maximize their dot product across all permutations of their components is to pair the components of each vector in increasing order. Similarly, to minimize their dot product one should arrange the first vector in increasing order and the second vector in decreasing order and take the dot product. For example, given $\mathbf{x} = (1, 2, 3, 4, 5)$ and $\mathbf{y} = (-1, -3, 7, -2, 0)$, it is easy to check that the maximum dot product achievable across all permutations is $7 \times 5 + 0 \times 4 + (-1 \times 3) + (-2 \times 2) + (-3 \times 1) = 25$ and the minimum dot product achievable is $7 \times 1 + 0 \times 2 + (-1 \times 3) + (-2 \times 4) + (-3 \times 5) = -19$.

Bounds on Estimators of θ

Theorem 2 (Bounds on General Estimator). *Let $\hat{\theta}_\omega$ be any estimator of θ, as defined in eq. (1). Suppose that the number of sites with j derived alleles is $\xi_j = k$, where $n - 1 \geqslant j \geqslant 1$. Then,*

$$kj\omega_j + (S_n - k) \max_{i \neq j}(i\omega_i) \geqslant \hat{\theta}_\omega \geqslant kj\omega_j + (S_n - k) \min_{i \neq j}(i\omega_i) \tag{3}$$

Equality with the upper bound is achieved only by the spectra $(0, \ldots, 0, S_n - k, 0, \ldots, 0, k, 0, \ldots, 0)$, where the position of $S_n - k$ is any coordinate $i \in \arg\max_{i \neq j}(i\omega_i)$ and the position of k is the jth coordinate; and equality with the lower bound is achieved only by the spectra $(0, \ldots, 0, S_n - k, 0, \ldots, 0, k, 0, \ldots, 0)$, where the position of $S_n - k$ is any coordinate $i \in \arg\min_{i \neq j}(i\omega_i)$ and the position of k is the jth coordinate. In particular, if $i\omega_i \neq i'\omega_{i'}$ for all indices $i \neq i'$, then the spectra achieving the upper and lower bounds are both unique.

Proof. First, let S_n be fixed. By fixing $x_j = k$, we see that $\boldsymbol{\xi} \in \{\mathbf{x} \in \mathbb{R}^{n-1} : x_i \geqslant 0 \ \forall i, x_j = k\} = \mathcal{R}$. We need to show that for any choice of $\boldsymbol{\xi} \in \mathcal{R}$, the dot product $(\omega_1, \ldots, (n-1)\omega_{n-1}) \cdot \boldsymbol{\xi}$ lies between $kj\omega_j + (S_n - k) \max_{i \neq j}(i\omega_i)$ and $kj\omega_j + (S_n - k) \min_{i \neq j}(i\omega_i)$. Hence, suppose $\boldsymbol{\xi} \in \mathcal{R}$ is given by $(\xi_1, \ldots, \xi_{j-1}, k, \xi_{j+1}, \ldots, \xi_{n-1})$, and we furthermore denote $i\omega_i = \tau_i$. Recall that in taking the dot product $\langle \boldsymbol{\tau}, \boldsymbol{\xi} \rangle = \tau_1 \xi_1 + \ldots + \tau_{n-1} \xi_{n-1}$, the term $\tau_j \xi_j = kj\omega_j$ is fixed; the τ_i are also fixed but the ξ_i $(i \neq j)$ are free. By Proposition 1,

$$kj\omega_j + \sum_{i \neq j} \tau_{[i]} \xi_{[i]} \geqslant \langle \boldsymbol{\tau}, \boldsymbol{\xi} \rangle \geqslant kj\omega_j + \sum_{i=1}^{n-1} \tau_{[i]} \xi_{[n-i+1]}.$$

In the expression above, $\xi_{[i]}$ denotes the ith largest component of $\boldsymbol{\xi}$ and $\tau_{[i]}$ denotes the ith largest component of $\boldsymbol{\tau}$. Finally, observe that the leftmost expression is at most $kj\omega_j + \tau_{[1]} \left(\sum_{i \neq j} \xi_{[i]} \right)$ — this can be shown by applying a simple weight shifting procedure. However, this last expression is precisely $kj\omega_j + (\max_{i \neq j} \tau_i)(S_n - k) = kj\omega_j + \max_{i \neq j}(i\omega_i)(S_n - k)$. Hence, the upper bound inequality in eq. (3) holds. By a similar argument, the lower bound inequality can also be derived. In case $i\omega_i \neq i'\omega_{i'}$ for all indices $i \neq i'$, the weight shifting procedure above must increase the value of the maximum dot product, unless $\xi_{[1]} = S_n - k$ in the first place. It thus follows that equality can only be achieved by the spectrum $(0, \ldots, 0, S_n - k, 0, \ldots, 0, k, 0, \ldots, 0)$, where the position of $S_n - k$ is the coordinate $i = \arg\max_{i \neq j}(i\omega_i)$ and the position of k is the jth coordinate. A similar argument establishes the equality condition with the lower bound in case $i\omega_i \neq i'\omega_{i'}$ for all indices $i \neq i'$. □

Theorem 2 implies constraints on the classical estimators of the neutral locus scaled mutation rate θ — which is equal to the average nucleotide diversity in cases where $N\mu$ is small — placed by fixing the value of ξ_j for some $j \in \{1, \ldots, n-1\}$. First, observe that to obtain the Watterson estimator $\theta_W = S_n/a_n$ we set $\omega_i = 1/ia_n$, so this implies that if S_n and n are known, then θ_W is fixed. However, for the rest of the estimators — notably Tajima's θ_π and Fay and Wu's θ_H, which places more weight on sites having higher derived allele counts — there is some "free room." We demonstrate how Theorem 2 provides explicit bounds on the amount of free room available.

Corollary 3 (Bounds on Tajima's θ_π)**.** *θ_π is obtained by the following choice of weights: for $n-1 \geqslant i \geqslant 1$, $\omega_i = (n-i)/\binom{n}{2}$. Suppose that the number of sites with j derived alleles is $\xi_j = k$, where $n-1 \geqslant j \geqslant 1$. Then,*

$$k\frac{j(n-j)}{\binom{n}{2}} + (S_n - k)\max_{i\neq j}\left(\frac{i(n-i)}{\binom{n}{2}}\right) \geqslant \theta_\pi \geqslant k\frac{j(n-j)}{\binom{n}{2}} + (S_n - k)\min_{i\neq j}\left(\frac{i(n-i)}{\binom{n}{2}}\right).$$

Moreover, equality with either the upper or the lower bound is achieved by the spectra given in Theorem 2.

Corollary 4 (Bounds on Fay and Wu's θ_H)**.** *θ_H is obtained by the following choice of weights: for $n-1 \geqslant i \geqslant 1$, $\omega_i = i/\binom{n}{2}$. Suppose that the number of sites with j derived alleles is $\xi_j = k$, where $n-1 \geqslant j \geqslant 1$. Then,*

$$k\frac{j^2}{\binom{n}{2}} + (S_n - k)\max_{i\neq j}\left(\frac{i^2}{\binom{n}{2}}\right) \geqslant \theta_H \geqslant k\frac{j^2}{\binom{n}{2}} + (S_n - k)\min_{i\neq j}\left(\frac{i^2}{\binom{n}{2}}\right).$$

Moreover, equality with either the upper or the lower bound is achieved uniquely by the spectra given in Theorem 2.

Corollaries 3 and 4 can be used to bound difference statistics in which one of the estimators is $\hat{\theta} = j\xi_j$. Two examples of such estimators (with $j = 1$) are given by Fu and Li in [7]: $F = \theta_\pi - \xi_1$ and $D = S_n/a_n - \xi_1$. Indeed, when one of the estimators is $j\xi_j$, it is likely that that the constraint placed by ξ_j on the other estimator reduces the range of values that the difference statistic can take. As it turns out, there is a general framework with which we can understand how generic tests for neutrality can be shown to be constrained by ξ_j.

Bounds on Difference Statistics

We prove a general result that establishes tight bounds placed by the number of sites with j derived alleles on values of any normalized difference statistic T_Ω.

Theorem 5 (Bounds on General Normalized Difference Statistic)**.** *Let T_Ω be any neutrality test under the infinite sites model, as defined in eq. (2). Suppose that the number of sites with j derived alleles is $\xi_j = k$, where $n-1 \geqslant j \geqslant 1$. Then,*

$$\frac{kj\Omega_j + (S_n - k)\max_{i\neq j}(i\Omega_i)}{\sqrt{\alpha_n S_n/a_n + \beta_n S_n(S_n-1)/(a_n^2 + A_n)}} \geqslant T_\Omega \geqslant \frac{kj\Omega_j + (S_n - k)\min_{i\neq j}(i\Omega_i)}{\sqrt{\alpha_n S_n/a_n + \beta_n S_n(S_n-1)/(a_n^2 + A_n)}}, \tag{4}$$

where $\alpha_n, \beta_n, a_n = \sum_{i=1}^{n-1} 1/i$ and $A_n = \sum_{i=1}^{n-1} 1/i^2$ as defined in the earlier Section. Moreover, if $i\Omega_i \neq i'\Omega_{i'}$ for all indices $i \neq i'$, then equality with the upper bound is achieved only by the spectrum $(0, \ldots, 0, S_n - k, 0, \ldots, 0, k, 0, \ldots, 0)$, where the position of $S_n - k$ is the coordinate $i = \arg\max_{i\neq j}(i\omega_i)$ and the position of k is the jth coordinate; and equality with the lower bound is achieved only by the spectrum $(0, \ldots, 0, S_n - k, 0, \ldots, 0, k, 0, \ldots, 0)$, where the position of $S_n - k$ is the coordinate $i = \arg\min_{i\neq j}(i\omega_i)$ and the position of k is the jth coordinate.

Proof. Similar to the proof of Theorem 2, the argument first uses the rearrangement inequality to bound the numerator, and then we divide both sides by the denominator which is fixed. □

Theorem 5 shows how values of neutrality tests can be constrained by additional knowledge of the number of polymorphic sites ξ_j satisfying any specified derived allele frequency j/n. We apply Theorem 5 to obtain specific tight bounds for classical tests for neutrality. To keep our presentation uniform, we continue using the symbol $T.$ to denote the test statistic instead of the usual letters (e.g., D and H). For example, we write Tajima's D as T_{Tajima}.

Corollary 6 (Bounds on Tajima's D)**.** *T_{Tajima} is obtained by the following choice of weights: $\Omega_i = (n-i)/\binom{n}{2} - 1/(ia_n)$, i.e., $i\Omega_i = i(n-i)/\binom{n}{2} - 1/a_n$. Suppose that the number of sites with j derived alleles is $\xi_j = k$, where $n-1 \geqslant j \geqslant 1$. Then,*

$$\frac{k\left(\frac{2}{n} - \frac{1}{a_n}\right) + (S_n - k)\left(\frac{n}{2(n-1)} - \frac{1}{a_n}\right)}{\sqrt{\alpha_n S_n/a_n + \beta_n S_n(S_n-1)/(a_n^2 + A_n)}} \geqslant T_{Tajima} \geqslant \frac{S_n\left(\frac{2}{n} - \frac{1}{a_n}\right)}{\sqrt{\alpha_n S_n/a_n + \beta_n S_n(S_n-1)/(a_n^2 + A_n)}}.$$

Moreover, equality with either the upper or the lower bound is achieved by the spectra given in Theorem 5.

Corollary 7 (Bounds on Fay and Wu's H)**.** *T_{FayWu} is obtained by the following choice of weights: $\Omega_i = (n-2i)/\binom{n}{2}$, i.e., $i\Omega_i = i(n-2i)/\binom{n}{2}$. Suppose that the number of sites with j derived alleles is $\xi_j = k$, where $n-1 \geqslant j \geqslant 1$. Then,*

$$\frac{\frac{k(n-2)}{n} + \frac{(S_n-k)n}{4(n-1)}}{\sqrt{\alpha_n S_n/a_n + \beta_n S_n(S_n-1)/(a_n^2 + A_n)}} \geqslant T_{FayWu} \geqslant \frac{\frac{k(n-2)}{n} + \frac{2(S_n-k)(2-n)}{n}}{\sqrt{\alpha_n S_n/a_n + \beta_n S_n(S_n-1)/(a_n^2 + A_n)}}.$$

Moreover, equality with either the upper or the lower bound is achieved by the spectra given in Theorem 5.

Application

Below, I discuss some potential applications of our results.

1. **Guiding the interpretation of and helping to explain values of estimators and T_Ω.**

   Underestimation of singletons in neutrality test based on NGS data due to ascertainment bias — including pooled samples and data for autopolyploids (e.g., plant genomes).

2. **Investigating the dynamics of estimators and T_Ω for data.**

   The trajectories of T_Ω and of $\hat{\theta}$ within the manifold bounding the pairs (ξ_j, T_Ω) and $(\xi_j, \hat{\theta})$ may lead to novel theoretical or empirical insights into the "stability" of statistics to evolutionary forces, or to perturbation of initial parameters.

References

[1] G. Achaz. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1):249–258, 2009.

[2] N. Alcala and N. A. Rosenberg. Mathematical constraints on fst: biallelic markers in arbitrarily many populations. *Genetics*, pages genetics–116, 2017.

[3] A. G. Clark, M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen. Ascertainment bias in studies of human genome-wide polymorphism. *Genome research*, 15(11):1496–1502, 2005.

[4] R. Durrett. *Probability Models for DNA Sequence Evolution*. Springer Science & Business Media, 2008.

[5] L. Ferretti, M. Perez-Enciso, and S. Ramos-Onsins. Optimal neutrality tests based on the frequency spectrum. *Genetics*, 186(1):353–365, 2010.

[6] L. Ferretti, S. E. Ramos-Onsins, and M. Pérez-Enciso. Population genomics from pool sequencing. *Molecular ecology*, 22(22):5561–5576, 2013.

[7] Y.-X. Fu and W.-H. Li. Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709, 1993.

[8] N. R. Garud and N. A. Rosenberg. Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps. *Theoretical Population Biology*, 102:94–101, 2015.

[9] M. Jakobsson, M. D. Edge, and N. A. Rosenberg. The relationship between F_{ST} and the frequency of the most frequent allele. *Genetics*, 193:515–528, 2013.

[10] T. S. Korneliussen, I. Moltke, A. Albrechtsen, and R. Nielsen. Calculation of tajima's d and other neutrality test statistics from low depth next-generation sequencing data. *BMC bioinformatics*, 14(1):289, 2013.

[11] A. W. Marshall, I. Olkin, and B. Arnold. *Inequalities: Theory of Majorization and Its Applications*. New York: Springer, 2010.

[12] S. B. Reddy and N. A. Rosenberg. Refining the relationship between homozygosity and the frequency of the most frequent allele. *Journal of Mathematical Biology*, 64:87–108, 2012.

[13] N. A. Rosenberg and M. Jakobsson. The relationship between homozygosity and the frequency of the most frequent allele. *Genetics*, 179:2027–2036, 2008.