Summer Statistics Research **Alan Aw**
Project Proposal alanaw1@stanford.edu
Supervisor: Julia Palacios Date: June 29, 2017
Co-supervisor: Noah Rosenberg

Aims and Objectives

In this project, we shall investigate theoretical properties of test statistics arising from the frequency spectrum, to derive mathematical relationships between values of these statistics and other features of the frequency spectrum. We shall also demonstrate the usefulness of such a mathematical treatment toward enhancing current statistical frameworks surrounding the effective use of such statistical tests on next generation sequencing (NGS) data, by means of data and mathematical plots that illustrate the mathematical relationships and their interaction with both sampled and simulated frequency spectra.

In addressing the questions raised above, we hope to enhance both existing and up-and-coming statistical frameworks developed for analyzing and interpreting NGS data, which should prove useful for both statisticians and practitioners applying the statistical methods to biological problems.

Background and Significance

In population genetics, statistics based on the frequency spectrum $\boldsymbol{\xi} = (\xi_i)_{i=1}^{n-1}$ often arise as unbiased estimators for parameters in evolutionary models. Such statistics are calculated on the genomic data and, at times, also on simulated spectra of a standard model. With the advent of NGS technologies, a rigorous statistical framework for obtaining optimal estimates of evolutionary parameters has also recently been formulated (see [6] and references therein). In computing these statistics, practitioners are able to (i) estimate evolutionary parameters (in case the parametric model is assumed correct), or to (ii) validate or reject the standard model in favor of an alternative evolutionary hypothesis. In both (i) and (ii), statistical problems abound—for (i), see, e.g., [4] for work on efficient algorithms for computing maximum likelihood mutation rates in a parametric inference framework. The focus of our project is on (ii).

First, consider a evolutionary model with mutation. In a population evolving neutrally, the population-scaled mutation rate (sometimes known as nucleotide diversity) $\theta = 4N_e\mu$ admits a family of unbiased estimators based on the frequency spectrum, which is given by

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

where $\boldsymbol{\omega} = (\omega_i)_{i=1}^{n-1} \in (\mathbb{R}_0^+)^{n-1}$ [2]. Neutrality tests based on the frequency spectrum compare two estimators of θ, where a test statistic called T_Ω is defined that is approximately $N(0,1)$ distributed, where

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

and $\hat{\theta}_{\omega_1}$ and $\hat{\theta}_{\omega_2}$ are any two estimators taking the form in eq. (1). Note that $\Omega_i \overset{\text{def}}{=} \omega_{1i} \Big/ \left(\sum_j \omega_{1j}\right) - \omega_{2i} \Big/ \left(\sum_j \omega_{2j}\right)$, and α_n, β_n are defined in eq. (9) of [2]. Classical tests of neutrality, including Tajima's D, Fu and Li's F and Fay and Wu's H, all correspond to specific choices of $\hat{\theta}_{\omega_1}$ and $\hat{\theta}_{\omega_2}$ in eq. (2) above. Under this framework, most powerful tests against a sufficiently large class of alternative evolutionary hypotheses can be obtained [5].

Despite the remarkable amount of theoretical work surrounding computational and statistical aspects of the neutrality test, little is known about the mathematical properties of the test statistic T_Ω. In similar mathematical studies of other population genomic statistics, including homozygosity-based tests for detecting sweeps (see [7]) and F_{ST} for measuring population differentiation (see [3] and references therein), it was shown that mathematical constraints on the test statistics can provide additional information about the genetic data. Hence, it remains relevant to investigate how mathematical properties of T_Ω may enhance interpretation of values computed of the test statistic.

Research Design and Methods

First, we will perform a thorough review of the literature on neutrality tests that involve T_Ω, to identify a reasonable scope within which to perform our investigation. Relevant papers are listed on the following GitHub page:

`https://github.com/JuliaPalacios/SummerResearch/blob/master/README.md`

Next, we will apply mathematical methods to obtain tight bounds on the relevant class of T_Ω to obtain a theoretical result. Finally, we will apply the theoretical bounds to simulated and sampled data, to compare how these data fit within the regions constrained by the bounds.

Attachments

Some preliminary results have been obtained, and are attached behind for reference.

References

[1] Achaz, Guillaume. (2008) “Testing for neutrality in samples with sequencing errors,” *Genetics* **179**: 1409-1424.

[2] Achaz, Guillaume. (2009) “Frequency spectrum neutrality tests: one for all and all for one,” *Genetics* **183**: 249-258.

[3] Alcala, Nicolas, Noah A. Rosenberg. (2017) “Mathematical constraints on F_{ST}: biallelic markers in arbitrarily many populations,” *Genetics*: in press.

[4] Bhaskar, Anand, Rachel Y.X. Wang, Yun S. Song. (2015) “Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data,” *Genome Research* **25**: 268-279.

[5] Ferretti, Luca, Miguel Pérez-Enciso, Sebastián Ramos-Onsins. (2010) “Optimal neutrality tests based on the frequency spectrum,” *Genetics* **186**: 353-365.

[6] Ferretti, Luca, Sebastián Ramos-Onsins, Miguel Pérez-Enciso. (2013) “Population genomics from pool sequencing,” *Molecular Ecology* **22**: 5561-5576.

[7] Garud, Nandita, Noah A. Rosenberg. (2015) “Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps,” *Theoretical Population Biology* **102**: 94-101.

Appendix

Below, I describe some preliminary results I obtained over this week.

1. *Main Result*

   Let the unfolded SFS of a given dataset ($n \geqslant 3$ sequences) be denoted by $\boldsymbol{\xi} = (\xi_i)_{i=1}^{n-1}$, where there are $\sum_{i=1}^{n-1} \xi_i = k$ segregating sites. Suppose the number of derived singletons $\xi_1 = \ell \in [0, k]$. Then, for any neutrality test $\Omega = \hat{\theta}_{\omega_1} - \hat{\theta}_{\omega_2} = \sum_{i=1}^{n-1} i\Omega_i\xi_i$ (where $\hat{\theta}_{\omega_1}$ and $\hat{\theta}_{\omega_2}$ are distinct estimators of θ),

   $$(k-\ell) \min_{2\leqslant i\leqslant n-1} (i\Omega_i) + \ell\Omega_1 \leqslant \Omega \leqslant (k-\ell) \max_{2\leqslant i\leqslant n-1} (i\Omega_i) + \ell\Omega_1.$$

   Moreover, equality in the left occurs if and only if $\boldsymbol{\xi} = (\ell, 0, \ldots, 0, k-\ell, 0, \ldots, 0)$ where $k-\ell$ occurs in the jth position of the vector and $j = \arg\min_{2\leqslant i\leqslant n-1}(i\Omega_i)$. Equality in the right occurs if and only if $\boldsymbol{\xi} = (\ell, 0, \ldots, 0, k-\ell, 0, \ldots, 0)$ where $k-\ell$ occurs in the jth position of the vector and $j = \arg\max_{2\leqslant i\leqslant n-1}(i\Omega_i)$.

2. *Specific Cases*

   - Tajima's $D = \Omega_{\text{Tajima}}$. Here, $\Omega_i = (n-i)/\binom{n}{2} - 1/(i\sum_{m=1}^{n-1} m^{-1})$, so

     $$i\Omega_i = \frac{(n-i)i}{\binom{n}{2}} - \frac{1}{\sum_{m=1}^{n-1} m^{-1}}.$$

     Hence,

     $$k\left(\frac{2}{n} - \frac{1}{\sum_{m=1}^{n-1} m^{-1}}\right) \leqslant \Omega_{\text{Tajima}} \leqslant \ell\left(\frac{2}{n} - \frac{1}{\sum_{m=1}^{n-1} m^{-1}}\right) + (k-\ell)\left(\frac{n}{2(n-1)} - \frac{1}{\sum_{m=1}^{n-1} m^{-1}}\right).$$

   - Fay and Wu's $H = \Omega_{\text{FayWu}}$. Here, $\Omega_i = (n-2i)/\binom{n}{2}$, so

     $$i\Omega_i = \frac{(n-2i)i}{\binom{n}{2}}.$$

     Hence,

     $$\frac{\ell(n-2)}{n} + \frac{2(k-\ell)(2-n)}{n} \leqslant \Omega_{\text{FayWu}} \leqslant \frac{\ell(n-2)}{n} + \frac{(k-\ell)n}{4(n-1)}.$$

   - Fu and Li's $F = \Omega_{\text{FuLi}}$. Here, $\Omega_i = \mathbf{1}(i=1)(2/n-1) + \mathbf{1}(i>1)\left[(n-i)/\binom{n}{2}\right]$, so

     $$i\Omega_i \overset{i\geqq 2}{=} \frac{i(n-i)}{\binom{n}{2}}.$$

     Hence,

     $$\frac{2(k-\ell)}{n} + \frac{\ell(2-n)}{n} \leqslant \Omega_{\text{FayWu}} \leqslant \frac{(k-\ell)n}{2(n-1)} + \frac{\ell(2-n)}{n}.$$

3. *Next Steps*

   We will obtain more bounds like the above, for the most powerful tests described in [5], and potentially even for the neutrality tests for NGS data which is discussed in [6]. Moreover, after this is done we will incorporate all bounds into the associated T_Ω statistics to obtain upper and lower bounds on T_Ω. The mathematical bounds restrict the region in which any computed value of T_Ω can lie. We will also explore ways of justifying the constraint set on the number of singletons ξ_1 (as opposed to other ξ_i's), potentially looking into arguments given in [1] regarding the omission of singletons in computing the test statistic for reasons owing to sequencing error underestimating the total number of singletons. Lastly, we will apply our bounds onto both sampled and simulated data sets.