Notes 20 : Tests of neutrality

MATH 833 - Fall 2012 *Lecturer: Sebastien Roch*

References: [Dur08, Chapter 2].

Recall:

THM 20.1 (Watterson's estimator) *The estimator*

$$\theta_W = \frac{S_n}{h_n},$$

is unbiased for θ. Its variance is

$$\mathrm{Var}[\theta_W] = \theta \frac{1}{h_n} + \theta^2 \frac{g_n}{h_n^2},$$

which converges to 0.

Also we will need the following result about the structure of the coalescent:

THM 20.2 *Assume that Π_i^n has sets of size $\lambda_1, \ldots, \lambda_i$ where the sets are ordered such that the first one contains 1, the second contains the smallest remaining element, etc. Let π be a permutation on $\{1, \ldots, i\}$ and define $\mu_\ell = \lambda_{\pi(\ell)}$, for $\ell = 1, \ldots, i$. Then the vector $(\mu_1, \ldots, \mu_i)$ is distributed uniformly over vectors summing to n.*

1 A class of θ estimators

A standard way to test whether the data is consistent with our model is to look at the difference between two different θ estimators. Most estimators in the literature can be expressed as follows. Let η_k be the number of segregating sites where the mutated alleles has frequency k, for $k = 1, \ldots, n-1$. This is known as the *site frequency spectrum*. (One can also define a similar notion when the ancestral states are not known, using the least common allele as a reference—leading to

the *folded site frequency spectrum*. See [Dur08, Section 1.4].) For constants $c_{n,k}$, $k = 1, \ldots, n-1$, consider the estimator

$$\hat{\theta} = \sum_{k=1}^{n-1} c_{n,k}\eta_k.$$

For instance, taking $c_{n,k} = 1/h_n$ for all k gives θ_W. (Recall that $h_n = \sum_{i=1}^{n-1} 1/i$.) Other important examples include:

- Choosing $c_{n,k} = 1$ if $k = 1$ and 0 otherwise gives $\theta_{FL} = \eta_1$.
- Choosing
$$c_{n,k} = \frac{k(n-k)}{\binom{n}{2}},$$
leads to θ_π, the pairwise differences.

To see that these are unbiased and potentially derive more estimators, we compute:

THM 20.3 *We have*

$$\mathbb{E}[\eta_k] = \frac{\theta}{k}.$$

For instance

$$\mathbb{E}[\theta_\pi] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} i = 1.$$

Proof:(Theorem 20.3) Imagine that at each level of the coalescent, the individuals are numbered uniformly at random. Let J_ℓ^k be the number of descendants of edge ℓ on level k. Letting L_m be the total branch length with m sampled descendants, note that

$$L_m = \sum_{k=2}^{n-m+1} t_k \sum_{\ell=1}^{k} \mathbb{1}\{J_\ell^k = m\},$$

where t_k is the time duration of level k. (Note that since each edge on level k has at least one sampled descendant, k must be smaller or equal to $n-m+1$ for ℓ to possibly have m sampled descendants.) So

$$\mathbb{E}[L_m] = \sum_{k=2}^{n-m+1} \frac{2}{k(k-1)} k\mathbb{P}[J_1^k = m].$$

By Theorem 20.2,

$$\mathbb{P}[J_1^k = m] = \frac{\binom{n-m-1}{k-2}}{\binom{n-1}{k-1}},$$

that is, the number of ways $k-1$ numbers sum to $n-m$ divided by the number of ways k numbers sum to n. Hence,

$$\begin{aligned}
\mathbb{E}[L_m] &= \sum_{k=2}^{n-m+1} \frac{2}{k(k-1)} k \frac{\binom{n-m-1}{k-2}}{\binom{n-1}{k-1}} \\
&= 2\frac{(n-m-1)!}{(n-1)!} \sum_{k=2}^{n-m+1} \frac{(n-k)!}{(n-m-k+1)!} \\
&= 2\frac{(n-m-1)!(m-1)!}{(n-1)!} \sum_{k=2}^{n-m+1} \binom{n-k}{m-1} \\
&= 2\frac{(n-m-1)!(m-1)!}{(n-1)!} \binom{n-1}{m} \\
&= \frac{2}{m},
\end{aligned}$$

where we used that the sum on the third line counts the number of ways to pick i elements from $n-1$ when the smallest one has index $m-1$. ∎

To compute the variance of the difference statistics, we also need the covariance between η_i and η_j. This is done in [Dur08, Section 2.1]. However, since the distribution of the statistics is not known, the tests are performed based on simulations.

2 Tajima's D and related statistics

Two common difference statistics are $\theta_\pi - \theta_W$ and $\theta_W - \theta_{FL}$. The former (standardized) is known as *Tajima's D statistic*, the latter, as *Fu and Li's D*. Under a null hypothesis of neutral evolution in a homogeneously mixing population of constant size, both statistics should be roughly 0. (However, their variance does not go to 0 as $n \to \infty$ because, unlike θ_W, θ_π and θ_{FL} are no consistent. Since we are not interested in estimating θ, this will not matter here.) Any significant deviation indicates the possibility that our assumptions may not be satisfied. To understand the effect of selection or population structure/growth on the difference

statistics, we note that

$$\theta_\pi - \theta_W = \sum_{k=1}^{n-1} \left(\frac{k(n-k)}{\binom{n}{2}} - \frac{1}{h_n} \right) \eta_k$$

and

$$\theta_W - \theta_{FL} = \sum_{k=1}^{n-1} \left(\frac{1}{h_n} - \mathbb{1}\{k=1\} \right) \eta_k.$$

In both cases, the coefficient of η_k is

- negative for small enough k;
- positive for middle k's.

In other words, an excess of singletons will tend to produce negative values and an excess of middle frequencies will tend to produce positive values.

Typical examples of phenomena creating these effects are:

- *Population growth* tends to produce a coalescent with long pendant edges (star-shaped genealogy), and therefore, an excess of singletons.
- *Deleterious mutations* will tend to produce very low frequency alleles, and thefore, an excess of singletons.

and

- *Population isolation* will tend to produce a delayed deepest coalescence (chicken-legs genealogy), and therefore, to create an excess of middle frequencies.
- *Balancing selection* will tend to produce an excess of heterozygotes, and therefore, an excess of middle frequencies.

Further reading

The material in this section was taken from Chapter 2 of the excellent monograph [Dur08].

References

[Dur08] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.