Harvard-MIT Division of Health Sciences and Technology
HST.508: Quantitative Genomics, Fall 2005
Instructors: Leonid Mirny, Robert Berwick, Alvin Kho, Isaac Kohane

Calculating Tajima's D

Fumio Tajima introduced a statistic that is widely used to test the null hypothesis of mutation-drift equilibrium and constant population size. Tajima considered two statistics: the mean pairwise difference (π) and the number (S) of segregating sites. Under the null hypothesis, the expected values of these statistics are

$$\begin{aligned} E[\pi] &= \theta \\ E[S] &= a_1\theta \end{aligned}$$

where $\theta = 4Nu$, $2N$ is the haploid population size, u is the mutation rate per generation, and a_1 is defined below.

Under the null hypothesis, π and S/a_1 both estimate θ, so they should be roughly equal in value. If they are about equal in value, then we cannot reject the hypothesis. If they are very different, on the other hand, we reject the hypothesis. But how different is "very different"? The answer depends on how variable these two statistics are from sample to sample. Tajima obtained a formula for the sampling variance of these statistics, and defined D this way: Let

$$V = \mathrm{Var}[\pi - S/a_1]$$

denote the sampling variance of the difference between the two estimates. Then Tajima's D is

$$D = \frac{\pi - S/a_1}{\sqrt{V}}$$

It expresses the difference between the two estimates relative to their standard error.

If the difference between π and S/a_1 were normally distributed, then we could expect D to lie between –2 and 2 about 95% of the time. In fact, this difference is *not* normally distributed, but it is not too far off. We should be suspicious of values of D that are much outside the interval $[-2, 2]$.

Although D is simple in concept, it is tedious to calculate. We need three pieces of data: π, S, and the number, n, of DNA sequences in the sample. Given these data, calculate:

$$
\begin{aligned}
a_1 &= \sum_{i=1}^{n-1} 1/i \\
a_2 &= \sum_{i=1}^{n-1} 1/i^2 \\
b_1 &= \frac{n+1}{3(n-1)} \\
b_2 &= \frac{2(n^2+n+3)}{9n(n-1)} \\
c_1 &= b_1 - 1/a_1 \\
c_2 &= b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2} \\
e_1 &= c_1/a_1 \\
e_2 &= c_2/(a_1^2 + a_2)
\end{aligned}
$$

Then Tajima's D is

$$
D = \frac{\pi - S/a_1}{\sqrt{e_1 S + e_2 S(S-1)}} \tag{5.5}
$$

This formula is exactly as given by Tajima in his 1989 paper. Hartl gives a slightly different form on page 113 of his text. The two formulas differ because Hartl defines π and S slightly differently than Tajima. For Hartl, π is the mean *fraction* of nucleotide sites that differ between pairs of individuals. For Tajima, it is the mean *number* of sites that differ. Similarly, for Hartl S is the *fraction* of the sites that are segregating and for Tajima S is the *number* of segregating sites. You will get the same answer using either approach. I have stuck with Tajima's.

Example Consider the following data set:

```
subj0     ATAATAAAAA AATAATAAAA AAATAAAAAA AATAAAAAAA A
subj1     AAAAAAAATA AATAATAAAA AAATAAAAAA AAAAAAAAAA A
subj2     AAAATAAAAA TATAATAAAA AAATATAAAA AAAAAAAAAA A
subj3     AAAAAAAAAA AATAATAAAA AAATAAATAA ATAAAAAAAA A
subj4     AAAATAAAAA AAATATAAAA AAATAAAAAA AAAAAAAAAA A
```

```
subj5      AAAATAAAAA AAAAATAAAA AAAAAAAAAA AAAAATAAAA A
subj6      AAAAAATAAA AATAATAAAA AAATAAAAAA AAAAAAAAAA A
subj7      AAAAAAAAAA AAAAATAAAA AAATAAAAAA AAAAAAAAAT A
subj8      AAAAAAAAAA AAAAAAAAAA AAATAAAAAA AAAAAAAAAA A
subj9      AAAAAAAAAA AAAAATAAAA AAATAATAAA AAAAAAAAAA A
```

From this data, we can calculate

```
10 sequences, 41 sites
                          pi: 3.888889
          Segregating sites: 16/41
theta_hat[estimated from S]: 5.655772
                 Tajima's D: -1.446172
a1=2.828968 a2=1.539768 b1=0.407407 b2=0.279012
c1=0.053922 c2=0.047227 e1=0.019061 e2=0.004949
```

Example Lynn Jorde's lab has published a large sample of DNA sequences from the D-loop of the human mitochondrial genome. There are 630 sites. For the Asian sample, we get:

```
77 sequences, 630 sites
                          pi: 8.438483
          Segregating sites: 103/630
theta_hat[estimated from S]: 20.958331
                 Tajima's D: -2.021749
a1=4.914514 a2=1.631862 b1=0.342105 b2=0.228184
c1=0.138626 c2=0.086985 e1=0.028208 e2=0.003374
```

For the African sample:

```
72 sequences, 630 sites
                          pi: 15.339984
          Segregating sites: 88/630
theta_hat[estimated from S]: 18.155855
                 Tajima's D: -0.525801
a1=4.846921 a2=1.630948 b1=0.342723 b2=0.228612
c1=0.136406 c2=0.085989 e1=0.028143 e2=0.003423
```

In one case, D is strongly negative and in the other case weakly negative. How would you interpret these results?