Examples of estimating θ and Tajima's D

These examples refer to pages 62-63 of the textbook, where estimation of θ using counts of the numbers of polymorphic sites is discussed. We'll consider two different data sets, which will illustrate the two different scenarios we discussed in class.

Data Set 1:

	1	2	3	4	5	6	7	8	9	10
Sequence A	A	C	G	C	G	T	C	T	A	A
Sequence B	A	C	G	C	G	A	C	T	A	G
Sequence C	A	C	G	A	G	A	C	T	A	A
Sequence D	A	T	G	C	G	A	C	T	A	A

Data Set 2:

	1	2	3	4	5	6	7	8	9	10
Sequence A	A	C	G	C	G	T	C	T	A	A
Sequence B	A	C	G	C	G	T	C	T	A	A
Sequence C	A	T	G	A	G	A	C	T	A	G
Sequence D	A	T	G	A	G	A	C	T	A	G

The data are clearly fake, but here's how I came up with them. Both data sets involve four mutations along different trees. There is one mutation on the lineage leading to each of the tips. If you think hard about the data (and look back at the trees in Figure 2.19 on page 63), you can probably guess which tree was used to write down each data set.

The goal here is to compute $\hat{\pi}$ and $\hat{\theta}_W$, and their difference, and to compare the differences (answers on the next page!).

Data set 1: The first data set was generated under a model in which all sequences diverged from a common ancestor at about the same time and evolved independently since then. With one mutation on the branch leading to each of the sequences, we have four polymorphic sites, sites 2, 4, 6, and 10, and $S_n = 4$. We also have the following matrix of pairwise counts of the number of polymorphic sites (e.g., in comparing sequences A and B, there are two polymorphic sites, so the entry in the matrix corresponding to A and B is 2):

	A	B	C	D
A	-	2	2	2
B	-	-	2	2
C	-	-	-	2
D	-	-	-	-

So we have the following estimates:

$$\hat{\pi} = \frac{2}{n(n-1)} \sum_{i<j} \pi_{ij} = \frac{2}{4(3)}(2+2+2+2+2+2) = 2$$

$$\hat{\theta}_W = \frac{S_n}{a_n} = \frac{4}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}} = \frac{24}{11} = 2.182$$

The difference between the two estimates is

$$\hat{\pi} - \hat{\theta}_W = 2 - 2.182 < 0$$

Data set 2: The second data set was generated under a model in which the tree is symmetric, with lineages A and B on one side of the root and C and D on the other. The time for the A-B and C-D lineages to coalesce to the most recent common ancestor is much longer than the time for the pairs A and B, and C and D, to coalesce. Thus, the mutation on the side of the tree leading to sequence A is much more likely to occur before the A and B lineages split than after, and a mutation along this lineage will affect both A and B. The number of polymorphic sites, S_n, is still 4, as in the previous example, but the nucleotide configuration at the polymorphic sites is different (compare them). Thus, $\hat{\pi}$ will change, but $\hat{\theta}_W$ will stay the same. We have the following matrix of pairwise counts of the number of polymorphic sites:

	A	B	C	D
A	-	0	4	4
B	-	-	4	4
C	-	-	-	0
D	-	-	-	-

The estimate of $\hat{\pi}$ is

$$\hat{\pi} = \frac{2}{4(3)}(0+4+4+4+4+0) = \frac{32}{12} = 2.667$$

The difference between the two estimates is

$$\hat{\pi} - \hat{\theta}_W = 2.667 - 2.182 > 0$$

So, as discussed in class, one scenario leads to a negative difference, and one leads to a positive difference. It is helpful to note what has changed between the two scenarios, namely $\hat{\pi}$. In the first case, every polymorphic site has a change in only one sequence, and so when forming pairwise differences, all comparisons show a low level of divergence. In the second case, every polymorphic site splits the taxa into two groups, and comparisons within group have no changes, while comparisons between groups show many changes. Since there are more between group than within group comparisons, $\hat{\pi}$ is larger.
When comparing the estimators, it is important to note that $\hat{\pi}$ incorporates frequencies of polymorphism to some extent, while the frequencies information is ignored in $\hat{\theta}_W$, which looks at only the total number of *sites* that are polymorphic.

One small note: our textbook defines these statistics using numbers of polymorphic sites. I have also seen these calculated using the percentage of polymorphic sites, which just scales things differently.