"PRINCIPLES OF PHYLOGENETICS: ECOLOGY AND EVOLUTION"
Integrative Biology 200 Spring 2018
University of California, Berkeley D.D. Ackerly

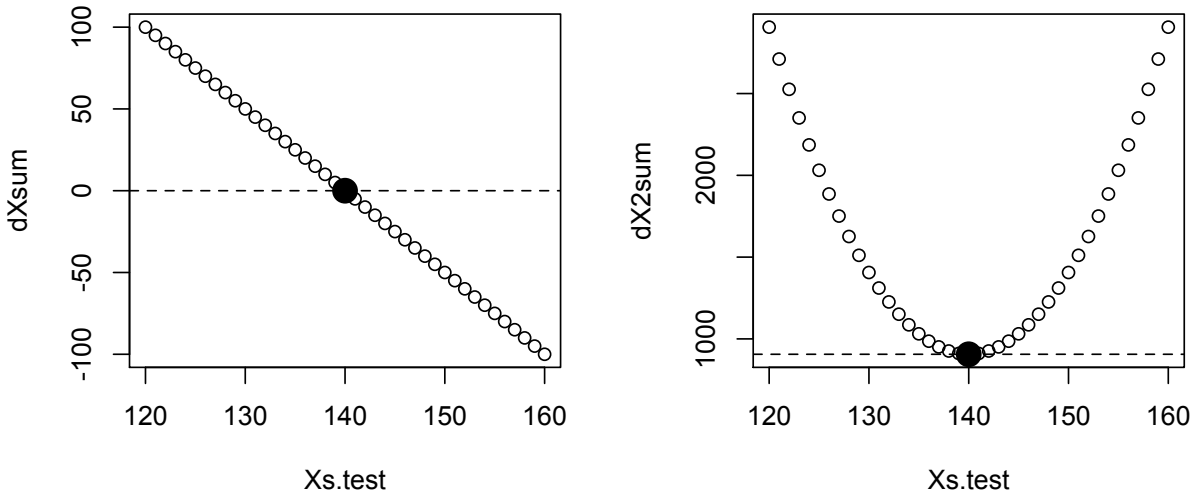
Feb. 21, 2018. **Sampling and inference**

Assigned reading: Tree Thinking: Beginning of Chapter 10 pp. 305 - 312

The goal of this lecture is to introduce and discuss the nature of inference in relation to standard statistical testing, and in particular the question of how we draw broad inferences from limited samples. Most of the discussion reflects generally on the nature of statistics, though some of the underlying philosophical issues have a different twist when we are making inferences about the past.

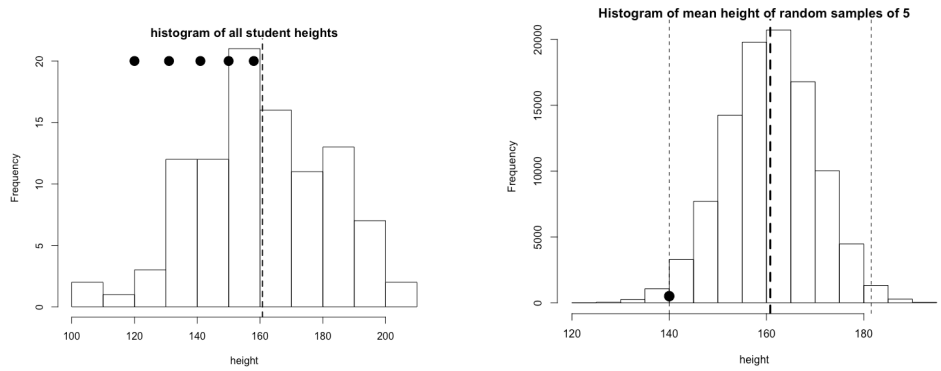
Descriptive and inferential statistics

- Let's pick 5 students in this room, and (for purposes of argument) say their heights are: 120, 131, 141, 150, 158 cm
- Mean = 140 cm
- What is a mean?
- Mean = sum(x)/N = value X* such that sum (X-X*) = 0
- Mean = value X* such that sum (X-X*)^2 is minimized



- Are the five students shorter (on average) than the average Berkeley student?
- Our 5 have a mean height of 140 cm
- Let's say that the average height of Berkeley students is 160 cm
- Is this group of people shorter than the average for Berkeley students? (Yes, No, Not enough information given)

- Rephrase our question: are the five students shorter than would be expected for a random sample drawn from the student population?
- We can think of the random sample being drawn from 1) all actual students or 2) all theoretically possible students, i.e. the general properties of the population of people who could be students at Berkeley



Left figure shows a hypothetical distribution of 100 students, with mean height of 160.78, sd = 21.17, and the heights of our five students. Now we have additional information about the scatter around that mean! Right figure shows the means for 100000 random draws of 5 students each from the distribution of 100 on the left. See how much more regular the histogram is? That's a feature of the central limit theorem – means calculated from samples drawn from any distribution tend to converge on a normal distribution.

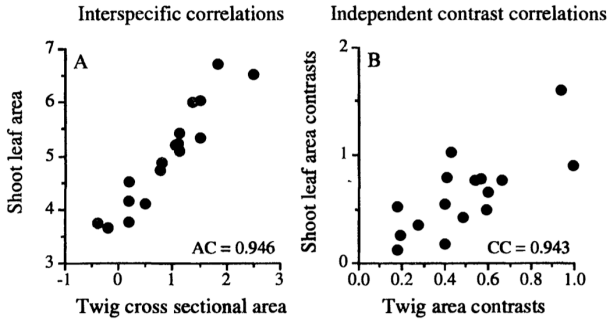
[Quick note and reminder: the standard error of a mean for a given N is the standard deviation you would expect if you took many samples of size N from the same underlying distribiution. E.g. above the full distribution has a sd = 21.17, while the standard deviation of the means on the right = 21.17/sqrt(5) = 9.47 (theoretical) and 9.24 (actual).]

Now we see that the sample of 5 with a mean of 140 seems to be quite unusual compared to random draws. In fact, only 1.34% of random samples have a mean that is lower than or equal to 140.

- Initial question: are the five students shorter than would be expected for a random sample drawn from the student population?
- Revised question: Can we reject the null hypothesis that these 5 students represent a random sample from the general population?
- Alternative hypothesis 1: students are non-random sample (2-tailed, p = 0.0268)
- Alternative hypothesis 2: students are non-random and shorter than average (1-tailed, p = 0.0134)
- We have decided as a sociological convention that we don't want to accept random patterns as evidence that something is actually going on more than 5% of the time!

In my post-doctoral research, I studied the evolution of canopy architecture in maples (*Acer*), integrating phylogenetic and trait data for a sample of 17 species, focused on those that regenerate in shade (from about 50 total in the clade) (Ackerly and Donoghue 1998). *Acer* is a clade of temperate trees distributed in Europe, East Asia and North America, nested in the largely tropical Sapindaceae. MRCA is about 50 Ma. They include understory and overstory trees, early and late successionals, and there is one very distinctive 'Japanese maple' clade of understory trees with shrubby, bifurcating branching architecture. I sampled species that regenerate in shade due to plasticity of canopy architecture in relation to light environment, and changes in functional significance of traits for species from shade vs. sunny environments.

Using independent contrasts (we'll learn more about that soon), I concluded that leaf area and twig size exhibit (exhibited?) positively correlated evolution in maples: as leaf area increases, so does twig size, or vice versa. Correlations do not imply a causal link, only that the two traits evolve together.



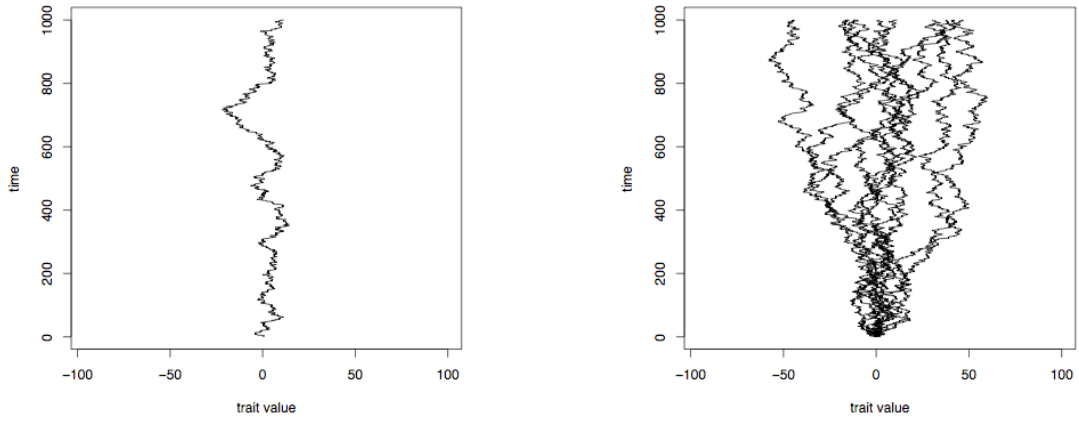
Why should this paper be publishable in a journal intended for a broad audience? One answer is methodological: here's how I did this work, you can do it too, for a group you are interested in. The other possibility is that we believe that these results are representative of what might be observed in other groups, so they are of interest to a broader audience interested in evolution and/or morphology, even if not interested in maples. Based on the information you've been given, which of the following inferences do you think is valid. Check all that apply:

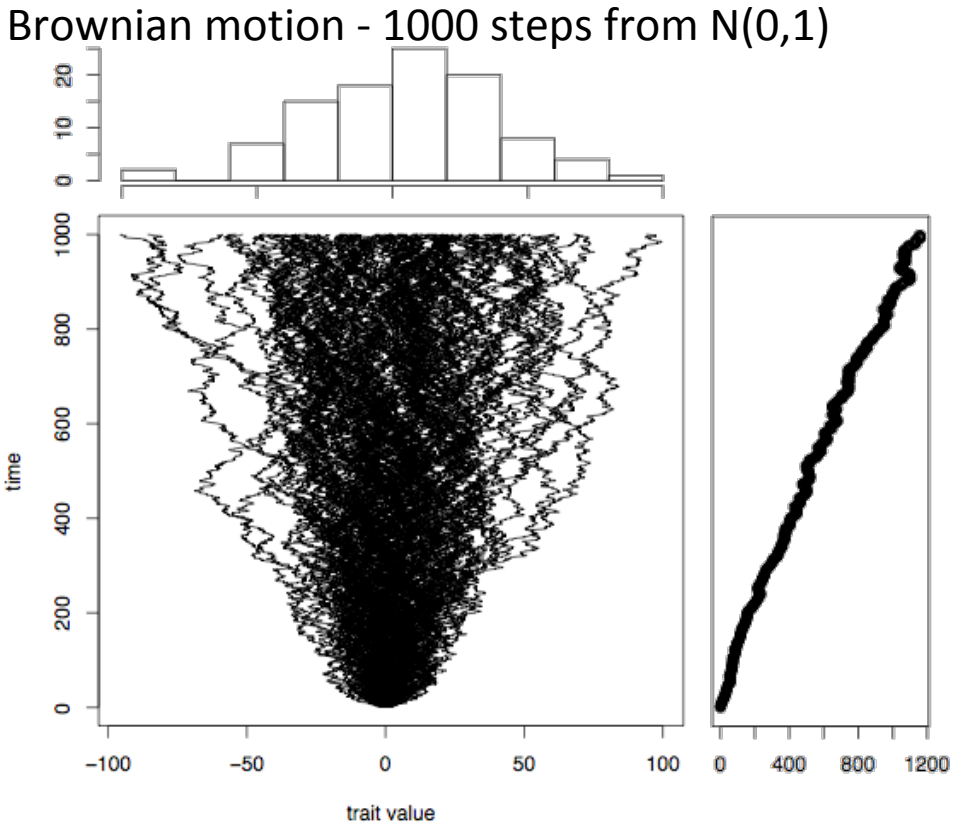
Leaf area and twig size coevolved in the evolution of:

- The 17 *Acer* in this study
- The evolution of *Acer* as a clade
- The evolution of temperate clades nested in tropical groups
- The evolution of Sapindaceae
- The evolution of shade-regenerating temperate trees
- The evolution of temperate trees
- The evolution of woody eudicots
- The evolution of woody plants
- The evolution of all plants
- The evolution of all organisms

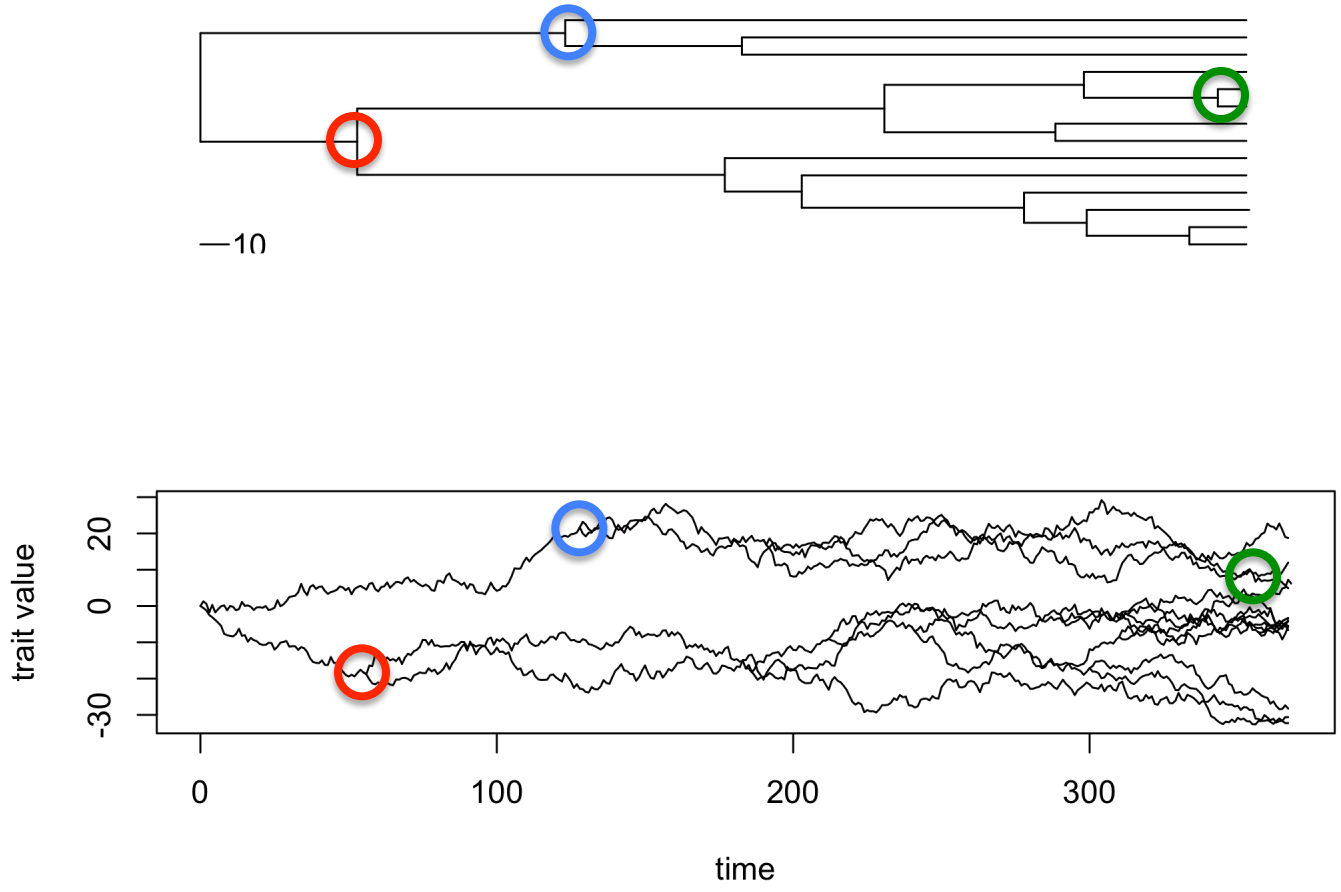
What is Brownian Motion?

- Random walk, non-directional
- Mirrors drift in population genetics, but does that mean it's only applicable to evolution by drift?
- When you have a model for the process that generates data, a single data point can have a variance – the expected variance if you were to generate that value repeatedly
- And the trait values for two species connected on a phylogeny can have a covariance – the expected covariance based on their degree of relatedness

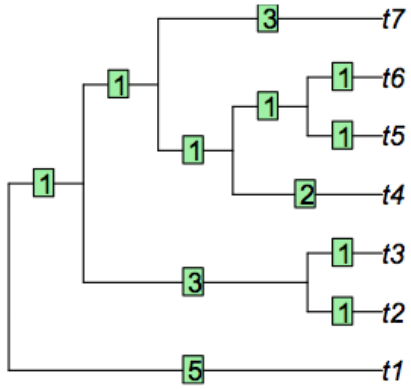




Brownian motion on a phylogeny



Traits of related species will be similar in proportion to the fraction of their history that's shared



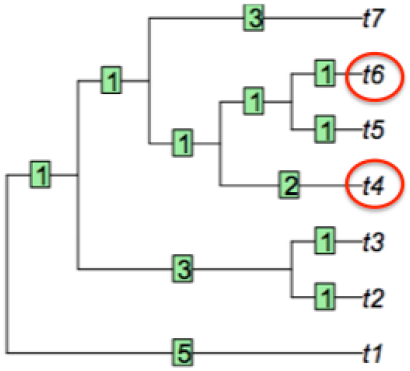
D - distance matrix

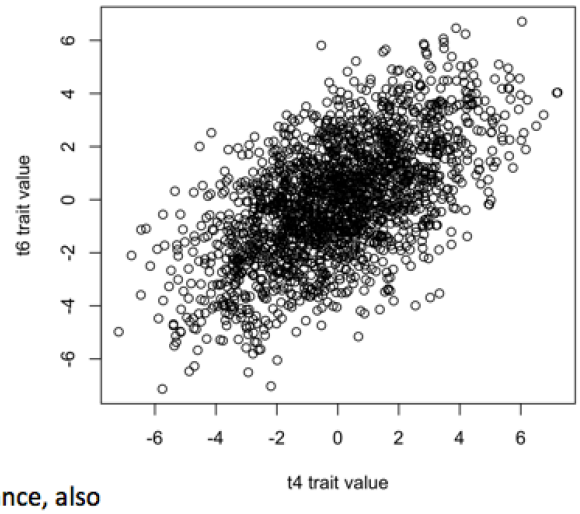
	t1	t2	t3	t4	t5	t6	t7
t1	0	10	10	10	10	10	10
t2	10	0	2	8	8	8	8
t3	10	2	0	8	8	8	8
t4	10	8	8	0	4	4	6
t5	10	8	8	4	0	2	6
t6	10	8	8	4	2	0	6
t7	10	8	8	6	6	6	0

C phylogenetic covariance matrix

	t1	t2	t3	t4	t5	t6	t7
t1	1	0.0	0.0	0.0	0.0	0.0	0.0
t2	0	1.0	0.8	0.2	0.2	0.2	0.2
t3	0	0.8	1.0	0.2	0.2	0.2	0.2
t4	0	0.2	0.2	1.0	0.6	0.6	0.4
t5	0	0.2	0.2	0.6	1.0	0.8	0.4
t6	0	0.2	0.2	0.6	0.8	1.0	0.4
t7	0	0.2	0.2	0.4	0.4	0.4	1.0

2000 reps, brownian motion
t4 vs. t6 values





trait values have an expected covariance, also called 'non-independence of species'