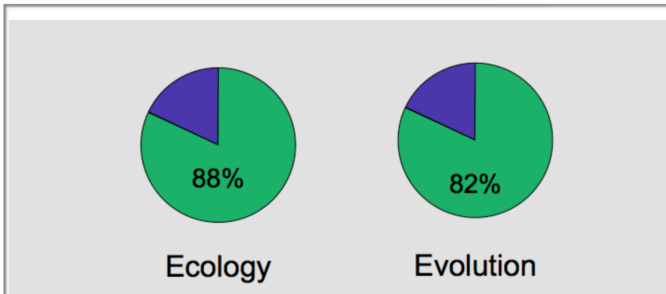
Welcome to StatsTree!

This website is designed to provide upper-level undergraduate students and graduate students will an understanding of biological statistics with a particular application to Ecology and Evolutionary Biology.

The Importance of Statistics to Ecology and Evolution

If you pick up a recent issue of a journal in your field you are likely to encounter some form of statistics in about 90% of the articles in that issue. The reason is that variation is present everywhere in the natural world, whether it is in morphology, behaviour, physiology, or ecology. This variation persists despite efforts by some of us to carefully control external sources of variation in a laboratory experiment. For other researchers collecting observational data, there can sometimes be phenomenal amounts of variation in the things they are interested in measuring.

Statistics is all about this variation. Without any variation there would be no need for statistics. There would also be no need for most of us as biologists since most of us are fundamentally interested in why biological entities differ.



Percentage of articles containing statistics in sample issues of the journals *Ecology* and *Evolution.*

Statistics is about the 'Signal to Noise Ratio'

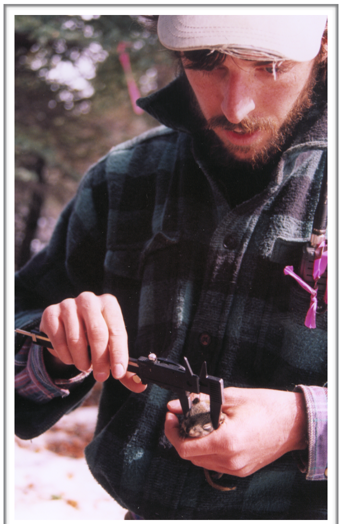
When we are trying to assess a relationship in statistics we are essentially weighing the magnitude of the effect against the amount of variation that is unexplained by the effect. Most test statistics (r in a correlation, t in a t-test, F in an ANOVA) are essentially just measures of the amount of signal (effect size) relative to the amount of noise (unexplained variation). As a result, it is important to think a bit about what causes variation.

There are two sources of variation that are really quite boring. The first is simply measurement error. If you are sloppy in collecting your data or if the measurement you are trying to take is particularly tricky to do (e.g. the animal squirms a lot) then repeated measurements of the same thing will differ just due to measurement error. This is not a biological phenomenon and is generally not of interest to us. This component of variation is important to statistics though because it does enter into the 'noise' part of statistics. So all I will say at this point is that if you are trying to identify a signal you ought to work hard to reduce your measurement error so as to minimize the amount of noise - variation in your data that is caused by you. The second source of variation is simple random variation.

Sources of Variation:

1. **Measurement error**
2. **Random variation**
3. **Complexity**

This is perhaps the most boring source of variation because it is usually not of biological interest to us and we aren't able to minimize it in any way. I mention it here mostly because what people generally attribute to 'random variation' is actually biological complexity, which is the final source of variation.



It is frequently under-appreciated that deterministic processes can lead to variability. Those of you that are familiar with quantitative genetics will know that one of Fisher's insights that was crucial to the modern synthesis in evolutionary biology was that the additive effects of many *deterministic* Mendelian loci could produce a continuously varying trait, such as size. [Aside: despite his contributions to evolutionary biology, Fisher was also a racist and should not be celebrated as a person.] So some of the variation in body size is due to the specific allele at locus A and the specific allele at locus B, etc. If we were sophisticated enough to be able to measure differences at all loci contributing to size then we would account for much of this variation. The same is of course true with respect to environmental causes of variation. Individuals might differ in size because of the amount of food they ate at a certain age, hormone levels during a critical period, light levels (if they are a plant), or perhaps temperature (especially for ectotherms). If you had information on all of these things we could potentially account for all of these sources of variation - both genetic and environmental.

So when you are trying to detect an effect in a variable measurement some of the variation in that measure unrelated to the effect that you are interested in will be due to biological complexity. This variation will work against your ability to detect your effect, so the more we can do to try and account for that complexity the greater our ability to detect our effect. The bottom line is that in any statistical test there is unexplained variation. It is important to remember that not all of this is due to measurement error or truly random variation. Much of it is due to biological complexity that we don't fully understand yet and the more we can do to sort out that complexity the more powerful our statistical test will be.

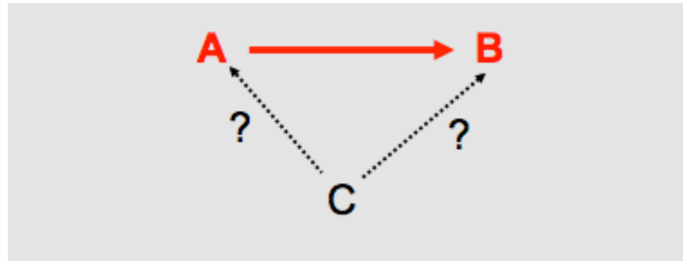
Approaches for detecting effects in the face of variability

There are several reasonable approaches to science (e.g. deduction versus induction) that I will not cover here. I would, however, like to highlight an important continuum in quantitative approaches - the continuum between experimentation and observation.

Experimentation attempts to identify effects by *manipulating* the variable of interest and carefully *controlling* for all other potential sources of variation. For example, the effect of a particular drug on blood pressure is tested on mice that are genetically identical, are at the same reproductive stage, fed the exact same food and housed in identical cages under very controlled and consistent conditions. This approach emphasizes that observed associations are not necessarily causal (correlation does not equal causation). Only controlled manipulations can break apart potentially confounding associations and identify true mechanisms. Some would argue that identifying mechanisms and causation is the ultimate goal of science and many funding agencies (e.g. NSF) highly value the approach. Others, however, would argue that we focus too much on mechanism. Some things are inherently phenomenological and have no mechanism (e.g. gravity). Others argue that our goals

Experimentation **Observation**

should not be mechanism, but instead prediction. As long as we are able to reliably predict important patterns do we really need to understand exactly what caused them? Rob Peters was a strong advocate for this latter view (see his book *A Critique for Ecology*).



Does an association between A and B really mean that the effect of A on B is causal? Perhaps the association is cause by some other unmeasured variable C that affects both A and B.

I would add one final point. Just because something **can** be shown to have an effect does not mean that it **does**. For example, we might manipulate CO_2 in a highly controlled greenhouse experiment and identify that increased CO_2 leads to increased growth rates of a particular plant. Some might consider this an elegant test of the effects of CO_2 on plant growth

rates. But just because CO_2 **can** affect growth doesn't mean that it **does**. It is entirely possible that this plant typically lives in competitive environments where growth rates are limited by the availability of sunlight and never reach levels at which CO_2 availability becomes limiting.

So what do you think? Where do you think the most important science lies along this continuum from Observation to Experimentation?

Goals of this website:

1. Provide an overview of modern statistical techniques
2. Introduce you to R, which is a powerful software package for statistical analyses
3. Give you experience analyzing data and presenting results relevant to ecology and evolution

Not our goals:

- Memorize formulae
- Promote point and shoot stats
- Survey the multitude of statistical errors to scare you from attempting new or more complex statistical techniques

Website Structure

Each topic on this website will be covered in three sections:

1. Videos. The primary content will be delivered through a series of short instructional videos.

2. Additional Resources will include relevant sections in the Whitlock & Schluter book "Analysis of Biological Data" including the relevant parts of the e-book called Sapling. Additional Resources will also include StatsTree handouts, example code and analysis in R as well as links to other useful external resources.

3. Review Questions provide an opportunity for you to assess your knowledge of the material before perhaps moving on to the next topic.