Lecture 1: Introduction. Why Statistics?

This course addresses the analysis of oceanographic observations and, as they become more complex, ocean models. Much of the material involves statistical procedures, which may appear foreign to students raised on a diet of deterministic, mathematical problems. The purpose of this lecture is to explain why a statistical perspective is appropriate in oceanography.

Even if we accept the proposition that the ocean can be exactly described by deterministic equations, there are three reasons why a statistical approach is practical.

1. The ocean is complex, requiring the specification of many variables, in fact, many more than can be observed.

2. The ocean is nonlinear, so that groups of variables cannot be studied in isolation.

3. Ocean observations are not controlled; all variables change as the system evolves.

Example of reason 1

Suppose we could write Newton's law applied to molecules to provide a complete, deterministic description of the ocean,

$$m_i \frac{d^2 \vec{x}_i}{dt^2} = \vec{f}_i(\text{all molecules}) \tag{1}$$

where m_i is the mass of the molecule at $\vec{x}_i$, t is time, and $\vec{f}_i$ is the force on molecule i depending on all other molecules. Writing these equations would be a challenge, to say the least. To get an idea of the magnitude of the challenge, we calculate the number of molecules of water in the ocean, given that the mass of the ocean is 1.4×10^{24} g,

$$1.4 \times 10^{24}\text{g} \times \frac{1\text{mole}}{18\text{g}} \times \frac{6.0 \times 10^{23}\text{molecules}}{1\text{mole}} = 4.7 \times 10^{46}\text{molecules} \tag{2}$$

Putting aside the question of how we determine the force on each molecule, initial conditions would be the position and velocity in three dimensions of each molecule. A complete initial specification would then require 2.8×10^{47} numbers. That this is a practically large number, consider the memory in a typical laptop computer, which might have 32 GB RAM. If we were to store these numbers in single precision format requiring 4 bytes, we would need 3.5×10^{37} laptops.

It is impractical to study the ocean by following the motion of molecules, so the approach is to treat the ocean as a continuous medium. We define a volume that is large enough to contain many molecules, and then a continuum velocity would be

$$\vec{v} = \frac{\sum m_i \vec{v}_i}{\sum m_i} \tag{3}$$

where the sum is over the volume. This is a mass weighted average over the volume, and is thus a statistical quantity. The volume must be defined so that it has many particles, and so that the smallest motions in the ocean are much larger than the volume. This is the case for a volume of about 1 mm on a side. There are

$$1.4 \times 10^{24}\text{ g} \times \frac{1\text{cm}^3}{1\text{g}} \times \left(\frac{10\text{mm}}{1\text{ cm}}\right)^3 \times \frac{1\text{ volume}}{1\text{ mm}^3} = 1.4 \times 10^{27}\text{ volumes} \tag{4}$$

such volumes in the ocean, which is still a large number. Clearly, there are more variables than can practically be observed or specified, so the ocean is complex.

Example of reason 2

The equation governing the momentum of a viscous fluid is the Navier-Stokes equation

$$\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \nabla)\, \vec{v} = -\frac{1}{\rho}\nabla p - g\hat{z} + \nu \nabla^2 \vec{v}, \tag{5}$$

which can be derived by application of (3) (Salmon 1998). The terms on the right-hand side are pressure gradient, gravity, and viscous friction. We have already concluded that it is impractical to apply this equation to the small volume used to define the continuum velocity in (3). The way forward is to do some more averaging until we finally get to a problem with few enough variables to specify that we can use our computer to solve it. Imagine an average that separates the "large-scale" features that we plan to treat as deterministic from the "small-scale" stuff that we will consider random. Denoting the average as $\langle \cdot \rangle$, $\vec{v}$ is

$$\vec{v} = \langle \vec{v} \rangle + \vec{v}' \tag{6}$$

where

$$\langle \vec{v}' \rangle = 0, \tag{7}$$

which makes explicit the separation into "large-scale", and "small-scale" . In the following we assume incompressibility

$$\nabla \cdot \vec{v} = 0. \tag{8}$$

Applying the averaging operator demonstrates that both large and small scale flow obey (8). Now consider the lefthand side of (5) for one component of the velocity, $\vec{v} = (u, v, w)$:

$$\frac{\partial u}{\partial t} + \vec{v} \cdot \nabla u. \tag{9}$$

Applying the average to this expression, and assuming that partial derivatives and averaging can be done in any order, produces

$$\frac{\partial \langle u \rangle}{\partial t} + (\langle \vec{v} \rangle \cdot \nabla) \langle u \rangle + \nabla \cdot \langle \vec{v}' u' \rangle\,. \tag{10}$$

The small-scale velocity unfortunately appears in the expression for the time rate of change of large-scale velocity. So our planned separation is not as simple as we might have hoped. In fact, we need to know the statistics of the small-scale in order to describe the large-scale. This last term in the expression is called the Reynolds stress, and its specification is one of the great unsolved problems in fluid mechanics.

> *I am an old man now, and when I die and go to Heaven there are two matters on which I hope enlightenment. One is quantum electro-dynamics and the other is turbulence of fluids. About the former, I am really rather optimistic.* Sir Horace Lamb, 1932.

> *There is a physical problem that is common to many fields, that is very old, and that has not been solved. It is the analysis of turbulent fluids.* Richard Feynman, 1963.

The Reynolds stress appears because advection is nonlinear, making it so that the large-scale flow cannot be studied in isolation of the small-scale flow. In general, parts of a nonlinear system cannot be studied in isolation. This is in marked contrast to linear systems, where, for example, waves of a single frequency can be studied in isolation of waves of all other frequencies. Acoustic waves in the ocean approach the ideal of linearity. Most other processes in the ocean are intrinsically nonlinear. Processes in the ocean that we may not wish to address then affect what does interest us in ways that are difficult to predict, and may appear to us to be random. Thus, a statistical approach is reasonable.

In recent years, a new wrinkle has been introduced to Reynolds averaging, by assuming that the averaging operator is a filter that smooths out small-scale structure but might not have zero mean. This process is referred to as "coarse-graining".

Example of reason 3

An oceanographer attempting to observe the ocean is doing something much different from a physicist doing an experiment in a laboratory. The quality of the physicist's experiment is often determined by how well extraneous factors are controlled so that the desired variable is measured with high accuracy. The best experimentalists are distinguished by their skill in devising effective controls. In contrast, oceanographers control almost nothing of what goes on in the ocean. An oceanographer interested in studying the general circulation will encounter tides, internal waves, eddies, turbulence, etc. which will confound observations. These uncontrolled processes may sometimes be usefully thought of as random, and statistical approaches are useful.

Suppose we want to calculate the mean ocean surface temperature at the Scripps pier. We could get all the data over the past 100 years and calculate the mean, but is that really what is desired? Should we account for the four seasons, and have a mean value for each? What about El Niños, which have a clear effect on temperature? In the end, we have to define over what data we average, and then our mean will reflect that choice.

This example requires us to define what we intent to be "signal" (mean temperature), and what processes are "noise" (seasons, El Niños, internal waves, etc.). By defining an average, we make explicit the separation between signal and noise we will get. Finally, there are a number of statistical tools we can use once we have defined the average. Defining signal and noise, and determining an ensemble over which to average are challenging; doing this intelligently is what makes a great observational scientist. Learning the statistical tools is easier, and is the subject of this class.

Lorenz model (demo)

In 1963, Ed Lorenz devised a model that encapsulates some of the complex issues laid out so far. This simple, deterministic model has only 3 variables and 3 equations, but its evolution appears to have random elements:

$$\begin{aligned} dX/dt &= -\sigma X + \sigma Y \\ dY/dt &= \rho X - Y - XY \\ dZ/dt &= -\beta Z + XY \end{aligned} \tag{11}$$

We solve these equations using σ=10, ρ=30, β=8/3, from $t = 0$ to 100. Consider the plot of X against Z, a plot of the phase plane (Figure 1). As you watch the phase plane fill in as a function of time, you can see that it is essentially impossible to predict where the trajectory is going. X takes

on values in either one of the "butterfly wings" in a detailed pattern (Figure 2). The reason that this may appear random is that we don't see Y in this plot. As the set of equations is deterministic, given values of all three variables the trajectory is exactly determined. One feature of apparent randomness is the crossings of trajectories. Given only X and Z at these points, it is impossible to know which branch the trajectory will take.

The results of this profoundly shape what we do. The Lorenz equations underscore the complexity of the ocean, the impact of nonlinearities, and the ways in which variables change as a system evolves. In the realm of prediction and predictability, the Lorenz equations tell us that model results are sensitive to boundary conditions, is we want to make accurate predictions, we need well constrained estimates of our initial conditions, that small anomalies (whether noise or just unusual features) can have big impacts on simulations.

One effort in modern data science is to go through a process of equation discovery, using data to find the equations or the parameters that underlie observations. What happens if you apply this process to data generated from the Lorenz system? Pietro Verzelli has an interesting blog post that attempts this, with decidedly mixed success. See https://verzep.github.io/Learning-Lorenz/.

Pseudorandom numbers

There are many ways to generate "pseudorandom" numbers on a computer. Because the numbers are generated on a computer, which can only do reproducible, deterministic calculations, the numbers are not truly random, so the term pseudorandom is used. A good pseudorandom generator takes on all possible values between 0 and 1, and has a very long period before it repeats. The pseudorandom generator on matlab rand uses the "Mersenne twister" algorithm and has a period of $(2^{19937} - 1)/2$. A time series created by using rand to make 3892 values is shown in Figure 3.

Statistical terminology To get started, here's a quick review of some definitions. A **random variable** is a variable whose value is determined by a **random process**. If a process produces values that are not perfectly predictable from what is known, it is a random process. A poorly controlled experiment might be considered a random process. Any function of a random variable is also a random variable. A **realization** of the process produces one random value of the variable. A large collection of realizations produced under statistically identical conditions (the same deterministic parameters) is an **ensemble** of "identically prepared" observations. The act of flipping a coin is a random process. If we let $x = 1$ for "heads" and -1 for "tails", x is a random variable. A single flip produces a realization of x and an afternoon of flipping using the same technique would yield an ensemble of realizations.

A central concept is the **average** or **expected value** of a random variable. The average of x is

$$\langle x \rangle = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} x_n \tag{12}$$

where x_n is the value of x in the n^{th} realization. Throughout these notes the symbol $\langle \cdot \rangle$ will be reserved for the ideal average requiring an infinite number of realization.

It is helpful to think of $\langle \cdot \rangle$ as a linear operator which can be applied to random variables. It is a linear operator because

$$\langle X + Y \rangle = \langle Y + X \rangle = \langle X \rangle + \langle Y \rangle, \quad \langle aX \rangle = \langle Xa \rangle = a\langle X \rangle, \quad \langle XY \rangle = \langle YX \rangle \neq \langle X \rangle \langle Y \rangle \tag{13}$$

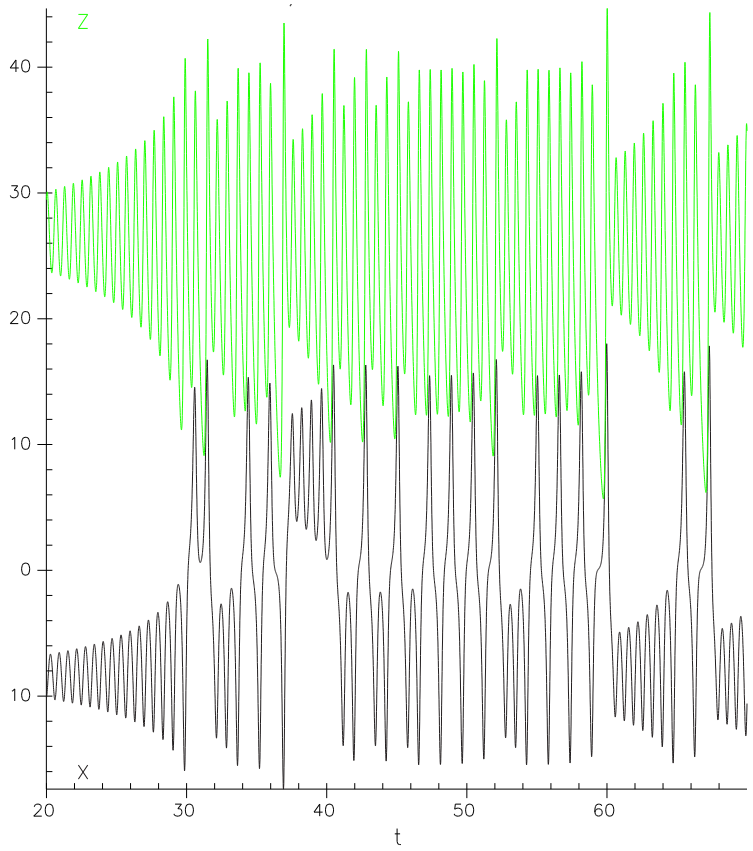

Figure 1: The variables X and Z from the initial evolution of the Lorenz model with $r = 28$. Both series start from $t = 0$ with initial conditions that differ by 0.1%. X shows that the evolution is completely different for $t > 31$. This is an example of a nonlinear, unpredictable system that is not complex (only 3 degrees of freedom).

when a is a constant.

Cumulative distributions and probability density functions

The complete description of a single random variable is its **distribution function** (or **cumulative distribution function**) D or, equivalently, its **probability density function** (pdf) F. These are defined by

$$D_x(r) \;=\; \text{Fraction of occurrences with } \; x < r, \tag{14}$$

where the fraction of occurrences is a probability. The notation might seem a little unusual: the random variable x is a subscript, and the distribution function depends on the deterministic variable r. Some properties of the distribution function are:

$$D_x(-\infty) \;=\; 0, \tag{15}$$

$$D_x(\infty) \;=\; \int_{-\infty}^{\infty} dr\; F(r) = 1, \tag{16}$$

$$D_x(r) \;\leq\; D_x(s) \text{ if } r \leq s. \tag{17}$$

The **probability density function** (pdf) is

$$F_x(r) \;=\; \frac{d}{dr} D_x(r) \quad \text{so that} \quad F_x(r)\, dr \;=\; \text{Fraction of occurrences with } \; r < x < r + dr. \tag{18}$$

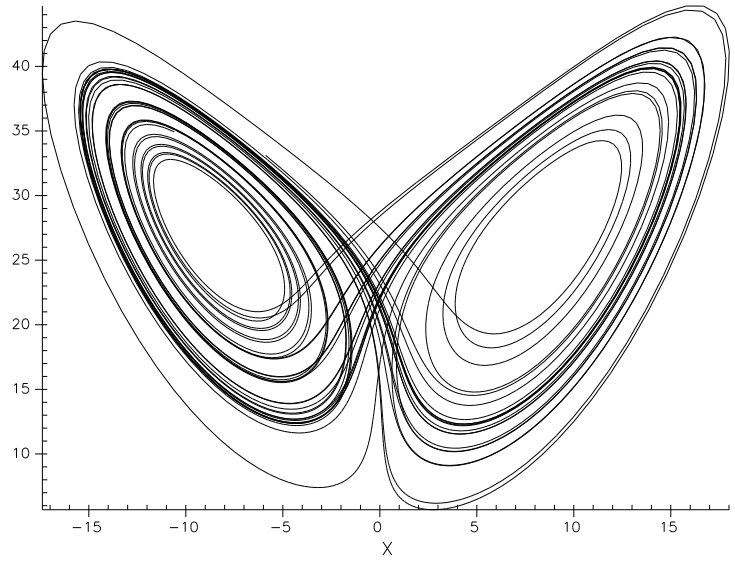

Figure 2: X-Z plane evolution of the Lorenz model shown in Figure 1. Data are taken from $t > 50$ when chaos has begun.

Some properties of the pdf are

$$F_x(r) \geq 0 \tag{19}$$

$$\int_{-\infty}^{\infty} F_x(r)\,dr = 1 \tag{20}$$

$$D_x(r) = \int_{-\infty}^{r} F_x(r)\,ds \tag{21}$$

Since the pdf is continuous, it is the limit of a histogram describing the number of occurrences in each of several "bins" of x, normalized so that the area under the curve is 1.

In the deterministic case that every realization produces the constant $x = A$, then $F_x(r) = \delta(r - A)$ and D would be a Heaviside unit step function with its edge at $r = A$. A handy representation of the probability density function is

$$F_x(r) = \langle \delta(r - x) \rangle. \tag{22}$$

To see why this is so, let N be an effectively infinite number of realizations of x and M be the number of realizations with $x < r$. Then, by definition,

$$D_x(r) \;=\; \frac{M}{N} \;=\; \frac{1}{N} \sum_{n=1}^{N} \int_{-\infty}^{r} dy\; \delta(y - x_n) \;=\; \int_{-\infty}^{r} dy \langle \delta(y - x) \rangle \tag{23}$$

from which (22) follows by differentiation with respect to r. The delta function is introduced into the third term to generate 1 when $x_n > r$ and 0 when it is not.

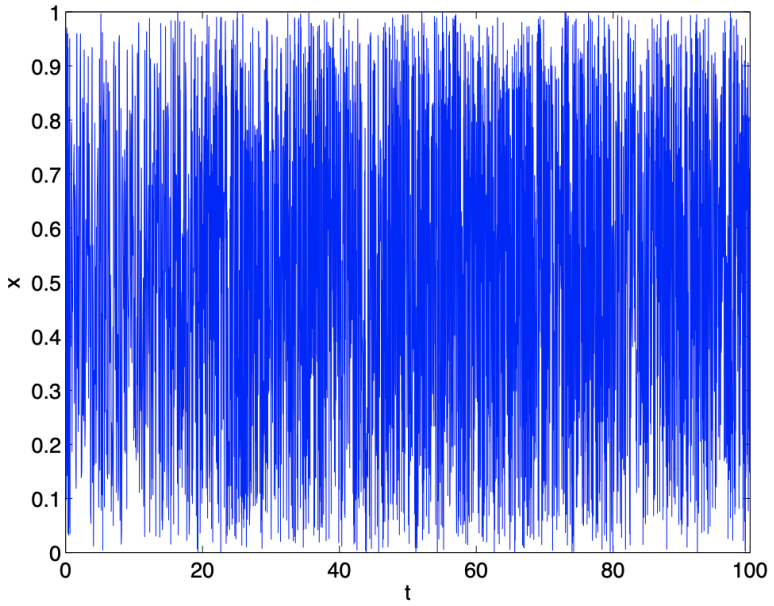

Figure 3: A series of uniformly distributed pseudorandom numbers between 0 and 1.