Why statistics?

Reading: Davis notes "WHY STATISTICS!"

Assume that all environmental variable are controlled by a large "deterministic" system. Such a system will have the following properties:

a) system is complex: more degrees of freedom than one can observe

⇓

the unobservable deg. of freedom

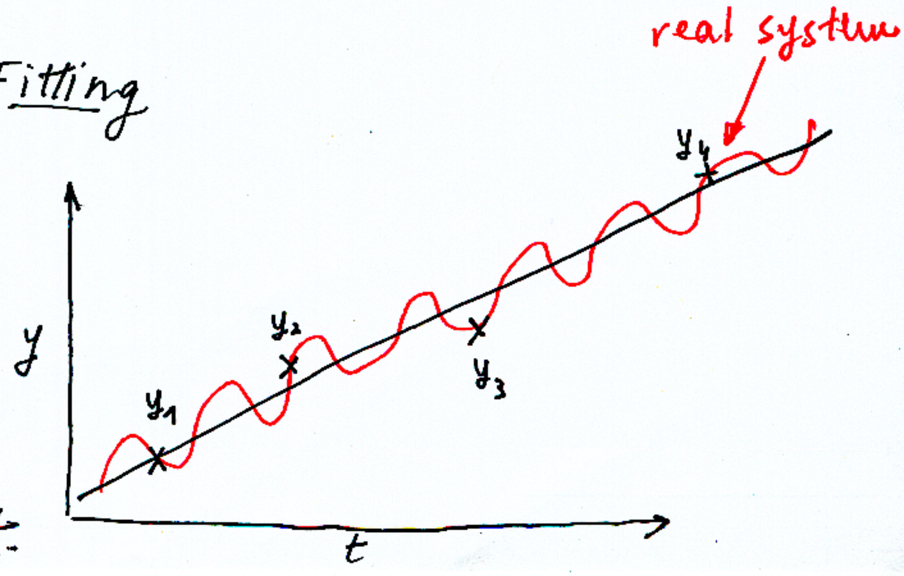
⇓

the system will appear to us as "non-deterministic" and introduce a "random" component

example → Function Fitting

$$y = at + \sin(\omega t + \vartheta)$$

If you only have 4 observation of y, the sin( ) component is going to appear like random noise.



b) system is ~~non~~ <u>non-linear</u>: variables cannot be studied in isolation.

1) e.g. El Niño, one cannot study ocean dynamics and atmosphere with out considering the coupling of the two systems.

c) dynamics are often <u>upredictable</u>:

it means that small changes in initial condition $O(\varepsilon)$ lead to order 1 $O(1)$ changes in the state of the system at future times.

NOTE: unstable linear system are also unpredictabe.

<u>CASE</u> of an underspecified and unpredictable system

$\Downarrow$

the unobserved degrees of freedom introduce a "random component" = "uncertainty"

$\Downarrow$

statistics are used to describe the typical behavior of a system when constrained by what is known.

An example of a system with few degrees of freedom FIGURE 1, which is also unpredictable

e.g.
- small changes in X initial conditions lead to drammatic differences in future state. Compare red and blue lines.
- variable Y studied in isolation appears to develop random fluctuations even though a time zero both the red and blue system have exact same state.

This simple deterministic system is chaotic, the governing dynamics are

$$\left.\begin{aligned} \frac{dX}{dt} &= -aX + aY \\ \frac{dY}{dt} &= rX - Y - XZ \\ \frac{dZ}{dt} &= -bZ + XY \end{aligned}\right\} \text{Lorenz Attractor}$$

A better description of the system is given by the "phase space diagram" FIGURE 2

- This shows the shape of the "attractor" in that the trajectory of the state in phase space collapsed around the attractor.

Assume you could not observe Y. Now multiple trajectory pass through the same point in $X-Z$ phase space. FIGURE 3, 4

$\Downarrow$

this apparent randomness arises from the missing or unobserved degrees of freedom

$\Downarrow$

in this context X and Z are "random signals" and Y is the "noise"

Let us analyze some basic statistics of X and Z_1

•) One fundamental statisitical quantity is the "Probability Density Function" (PDF)

FIGURE 5.

••) The PDF tells us the probability that our random signal, in this case X, Z_1, will have a certain value. Note that the sample mean is not necessarily the most likely value.

For this case the single PDF is not very helpfull.

•••) Let us combine the statistics of X and Z_1 to obtain a "Joint PDF": JPDF tells us the probability that X and Z_1 will have certain values

FIGURE 6

••••) Let us now assume that we know something about Y → conditional JPDF. FIGURE 7

- Most environmental system have multiple (many) degrees of freedom and description like the phase space and joint PDFs are not very useful (and adequately measured)

$\Rightarrow$ separate the system into:

signal + noise

signal: is what we hope to resolve!

noise: what we cannot fully resolve

$\Downarrow$ $\dfrac{\text{signal}}{\text{noise}}$ becomes and important ratio. If $\gg 1$ you have a good set of data

$\Downarrow$

derive useful statistics

> The aim of statistics is to deal with the essence of the process without dealing with it in detail

Statistical Data Analysis can be viewed as 3 steps:

1) separating signal and noise

2) defining the ensemble over which a typical behaviour can be defined

3) develop an accurate statistical description using the ensemble

1 and 2 are problematic.