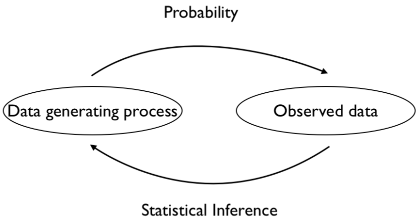
Class Overview

Elizabeth Purdom

August 14, 2024

The Big Picture



adapted from Wasserman, 2004

The Big Picture: Example

We consider a simple problem of visitors to a website. We could use probability to model this process.

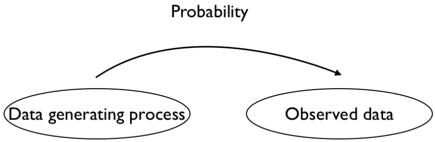
For example, we could assume that the number of visitors in any interval of time t is distributed $Poisson(\lambda|t)$ and independent of the number in any other interval of time s, where λ is the rate of visitation to the site

This would be the common (homogenous) Poisson process as a model for the distribution of the occurrence of visitors across time.

The Big Picture: Example

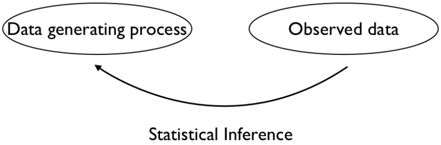
We could use this probabilistic model to detail

- on average how many visitors to expect in a particular time interval
- the chances of two visitors arriving within δ seconds of each other for any particular rate λ
- determine the λ at which the probability is greater than 0.80 that the demand in a time interval δ would be above capacity of the host (i.e. crashes).



The Big Picture: Example

Alternatively, you could have records of visits to the website for a short period of time, and you would like to determine what is the actual rate λ.



Notice I still assume a probabilistic model generated the data, but rather than trying to understand the implications of the type of data that the probability model will generate, I want to *estimate* something about that probability model. More than that – I'd like to be able to say something about how accurate my estimate is.

Basic setup

- We have a sample of data $X = (X_1, \ldots, X_n)$ from a distribution F (F can refer to the distribution or the CDF).

Basic setup

- We have a sample of data $X = (X_1, \ldots, X_n)$ from a distribution F (F can refer to the distribution or the CDF).
- A statistical model $\mathcal{F}$ is a set of possible distributions F that could have generated the data $X = (X_1, \ldots, X_n)$,

Basic setup

- We have a sample of data $X = (X_1, \dots, X_n)$ from a distribution F (F can refer to the distribution or the CDF).
- A statistical model $\mathcal{F}$ is a set of possible distributions F that could have generated the data $X = (X_1, \dots, X_n)$,
- Inference is the process of using the data X to draw conclusions about F

Basic setup

- We have a sample of data $X = (X_1, \dots, X_n)$ from a distribution F (F can refer to the distribution or the CDF).
- A statistical model $\mathcal{F}$ is a set of possible distributions F that could have generated the data $X = (X_1, \dots, X_n)$,
- Inference is the process of using the data X to draw conclusions about F
- Our conclusions might not be about the entire distribution, but some feature (or parameter) of the probability distribution θ that we are interested in.

Overview of Inference Procedures

Different classes of models $\mathcal{F}$:

- Nonparametric inference
- Parametric inference

Different modeling paradigms

- Frequentist inference
- Bayesian inference

Different types of inference:

- point estimation
- confidence sets (Frequentist) or credible regions (Bayesian)
- hypothesis testing (largely frequentist)

What is the point of this class?

- We will discuss mathematical properties of methods that try to answer these questions
- This class is mathematical in focus, but not great technicality.
  - Theorems will be given, generally not proved unless simple
  - You will be asked to derive results, but again not complicated or technically intricate.

What is the point of this class?

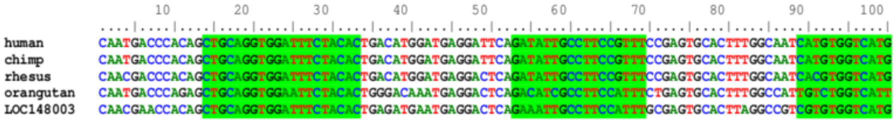
- We will discuss mathematical properties of methods that try to answer these questions
- This class is mathematical in focus, but not great technicality.
  - Theorems will be given, generally not proved unless simple
  - You will be asked to derive results, but again not complicated or technically intricate.
- Goal is not to necessarily learn methods – we will spend a lot of time talking about very simple examples!

What is the point of this class?

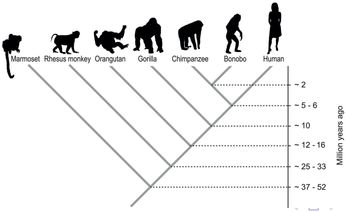
- We will discuss mathematical properties of methods that try to answer these questions
- This class is mathematical in focus, but not great technicality.
  - Theorems will be given, generally not proved unless simple
  - You will be asked to derive results, but again not complicated or technically intricate.
- Goal is not to necessarily learn methods – we will spend a lot of time talking about very simple examples!
- Idea is to learn how to think about methods – what makes a good method? how do you show it? – and basic strategies for developing methods.
- Basic building blocks that can be used as a starting point for much more complicated method development
- Learn the language for talking about these concepts

More Complicated Example: Phylogentic Trees

- We know that organisms evolve and that this evolution is done through a process of changes introduced into DNA that change the physical features of organisms.
- We have data in the form of DNA from organisms



- We want to understand the path of evolution that created the DNA we observe



More Complicated Example: Phylogentic Trees

- We can think of this phylogentic tree as a complicate parameter θ that we want to estimate

More Complicated Example: Phylogentic Trees

- We can think of this phylogentic tree as a complicate parameter θ that we want to estimate
- We can develop a probability model for how DNA is created if you know the tree.

More Complicated Example: Phylogentic Trees

- We can think of this phylogentic tree as a complicate parameter θ that we want to estimate
- We can develop a probability model for how DNA is created if you know the tree.
- Then to estimate the tree
  - Develop Maximum Likelihood method
  - Develop a Bayesian method
  - Use other (non-parametric) methods

More Complicated Example: Phylogentic Trees

- We can think of this phylogentic tree as a complicate parameter θ that we want to estimate
- We can develop a probability model for how DNA is created if you know the tree.
- Then to estimate the tree
    - Develop Maximum Likelihood method
    - Develop a Bayesian method
    - Use other (non-parametric) methods
- We might then want to
    - Give confidence on a node of the tree
    - Ask which of these methods is better
    - Consider performance under alternative probability models