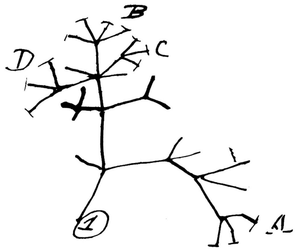


IPS-164 INTRODUCTION TO PHYLOGENETICS 2022

Lecture 6
Intro to statistical phylogenetics. Part I

Sergei Tarasov

Beetle curator & Docent
Finnish Museum of Natural History, University of Helsinki

- @tarasov_sergio
- sergei.tarasov@helsinki.fi
- https://www.tarasovlab.com

LET’S GET TO KNOW EACH OTHER

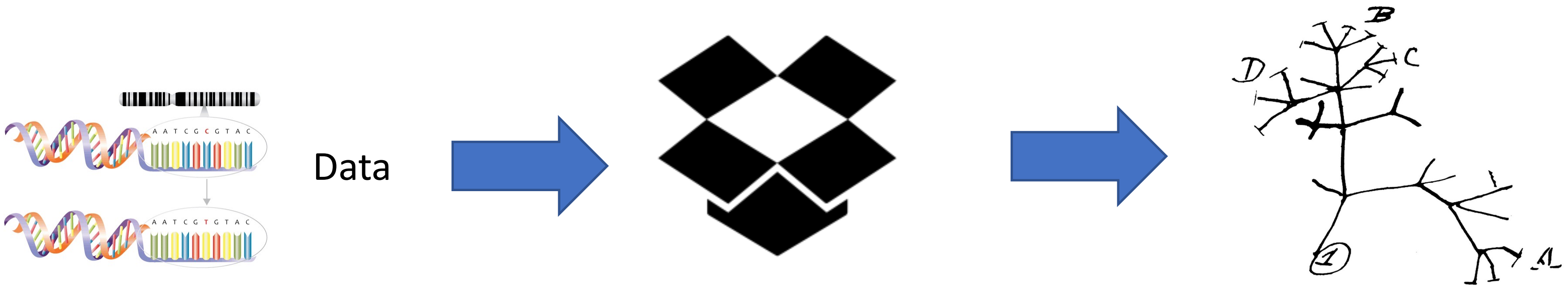
1. Say your name
2. Your interests in biology/phylogenetics
3. What has brought you to this course?
4. Your expectations from the course?

PLAN OF THE TODAY'S LECTURE

1. Intro to this (the second) part of the course
2. Overview of the statistical phylogenetics: which questions statistical phylogenetics can address?
3. Parsimony vs. statistical phylogenetics
4. Intro to statistics and modeling
5. Binomial model

Aim of this part of the course

- Explain how the statistical inference works in phylogenetics and overview its main field
- So, you will be able to calculate likelihood by “hand”
- You will learn how to reconstruct phylogenetic tree and perform various other analyses
- You will be able to select amongst available methods to address your research questions



Data

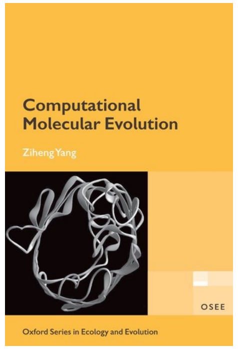
Inference

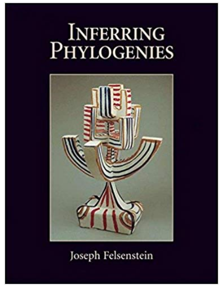
Lectures in this part of the course

6. Introduction to statistical phylogenetics (part I)
7. Introduction to statistical phylogenetics (part II)
8. Reconstructing phylogenies (part I)
9. Reconstructing phylogenies (part II)
10. Tree dating
11. Trait evolution
12. Trait evolution and Diversification

Suggested literature

- Computational Molecular evolution

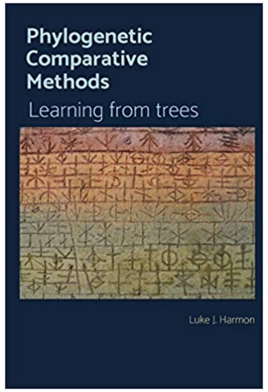

Computational Molecular Evolution
Ziheng Yang
OSEE
Oxford Series in Ecology and Evolution

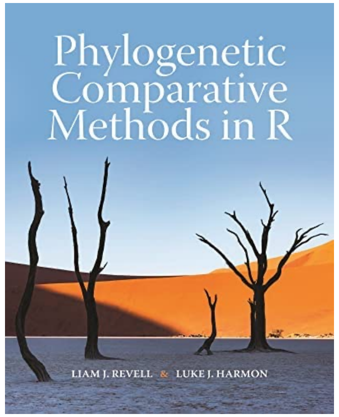


- Felsenstein's Inferring phylogenies

- Luke Harmon. Phylogenetic Comparative Methods

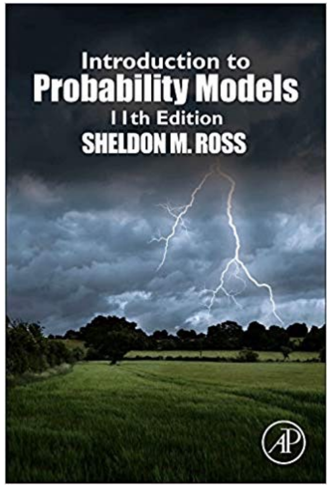
  It's free at https://lukejharmon.github.io/pcm/



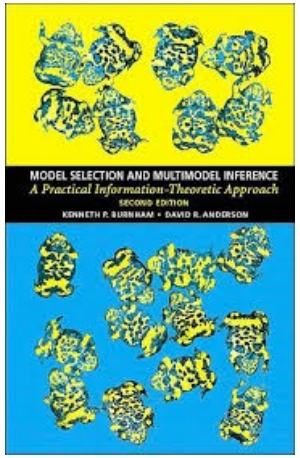

Phylogenetic Comparative Methods in R
LIAM J. REVELL & LUKE J. HARMON

- Phylogenetic Comparative Methods in R

Suggested literature



- Intro to Probability models
- Model Selection by Burham and Anderson


MODEL SELECTION AND MULTIMODEL INFERENCE
A Practical Information-Theoretic Approach
SECOND EDITION
KENNETH P. BURNHAM • DAVID R. ANDERSON

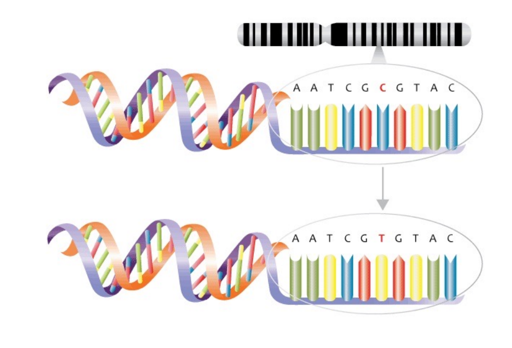
Developing skills in Statistical Programming with R or Python

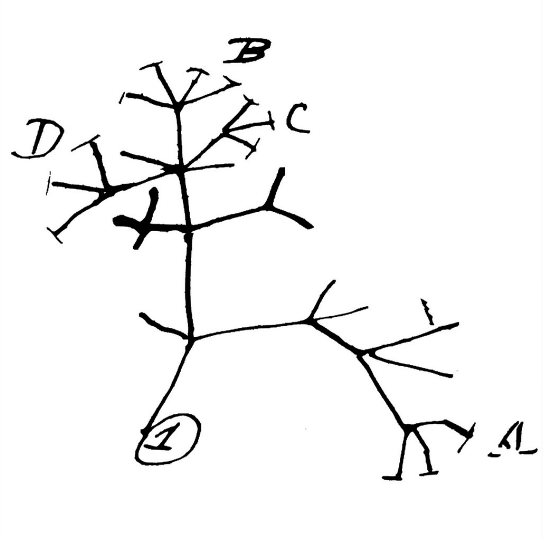
- Books
  - Paradis. Analysis of Phylogenetics and Evolution with R
  - Revell & Harmon. Phylogenetic Comparative Methods in R
  - Grolemund. R for Data Science. https://r4ds.had.co.nz/index.html

- All materials will be available at:
  - GitHub https://github.com/sergeitarasov/Course_IPS-164
  - UH website https://www.mv.helsinki.fi/home/jhyvonen/IPS-164/

WHICH QUESTIONS STATISTICAL PHYLOGENETICS CAN ADDRESS

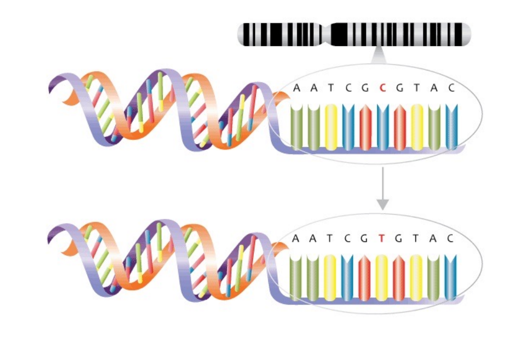
1. Tree Reconstruction

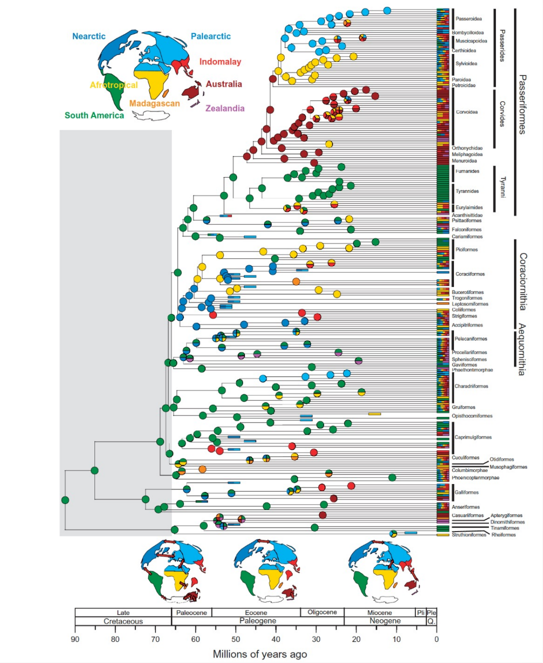




WHICH QUESTIONS STATISTICAL PHYLOGENETICS CAN ADDRESS

2. Tree Dating

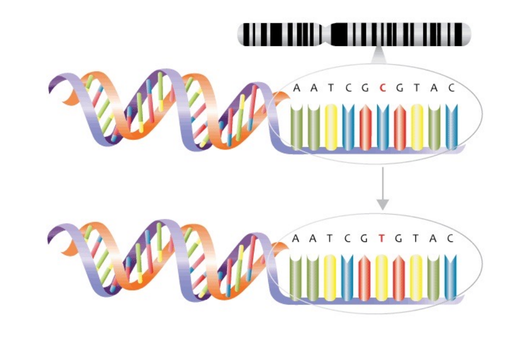


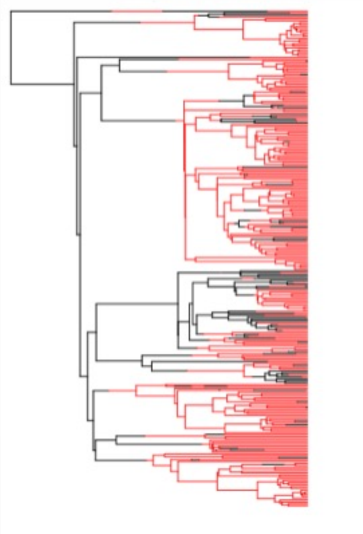


Fossil record

WHICH QUESTIONS STATISTICAL PHYLOGENETICS CAN ADDRESS

3. Ancestral Character State Estimation



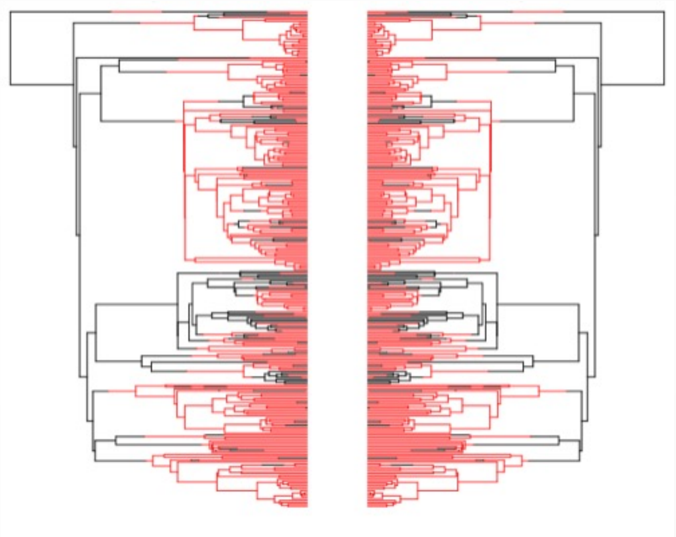


Discrete and Continuous traits

Give you example of traits!

WHICH QUESTIONS STATISTICAL PHYLOGENETICS CAN ADDRESS

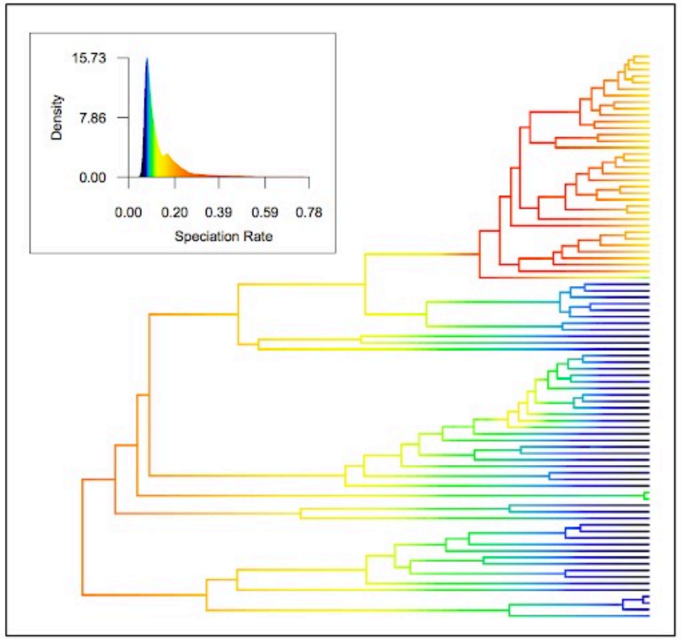
4. Correlation Between two or more traits



Discrete and
Continuous traits

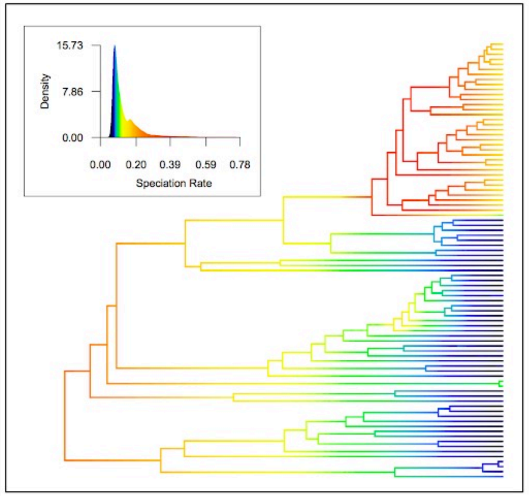
WHICH QUESTIONS STATISTICAL PHYLOGENETICS CAN ADDRESS

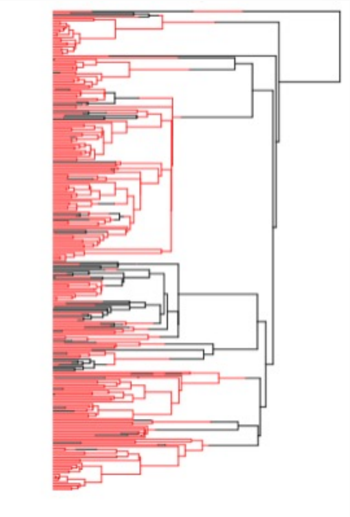
5. Reconstructing Diversification process



WHICH QUESTIONS STATISTICAL PHYLOGENETICS CAN ADDRESS

6. Correlation between Diversification and traits





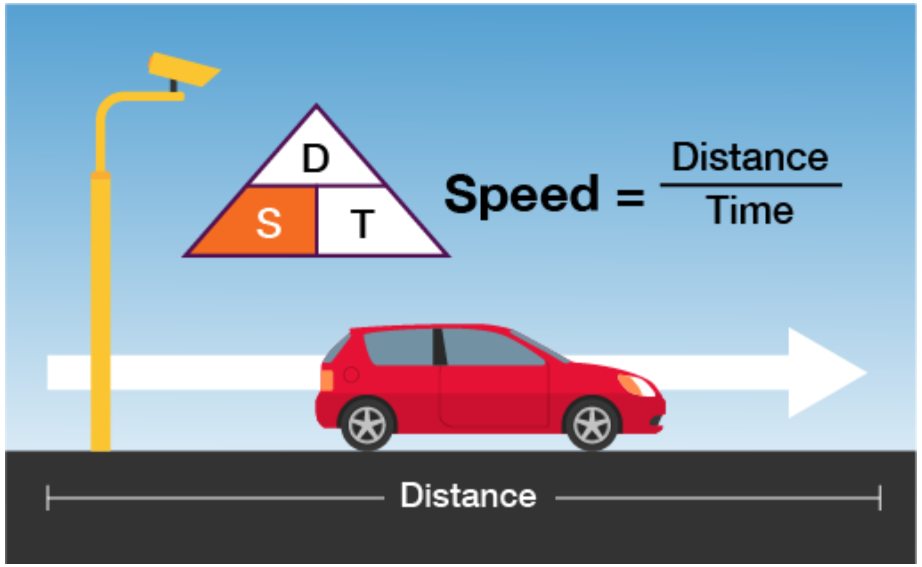
MAIN STATISTICAL METHODS

Likelihood and Bayesian inferences

Ronald Fisher

Thomas Bayes

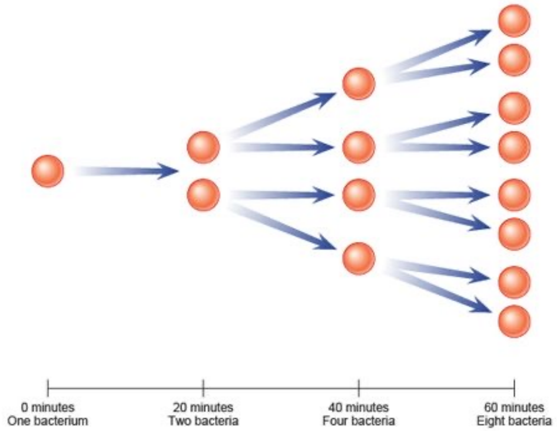
Modeling natural and phylogenetic phenomena



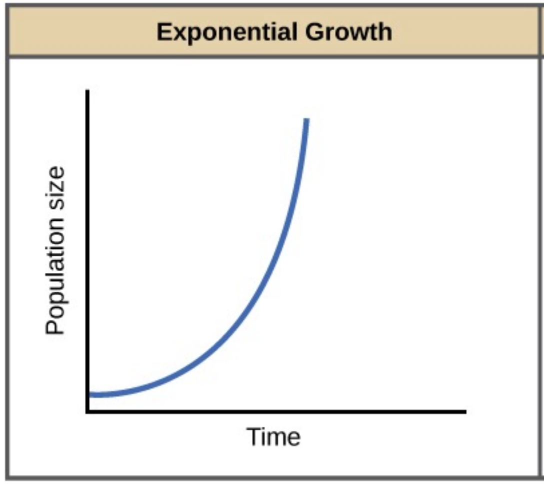
Models are common in physics

Models are sets of rules describing how a system changes over time

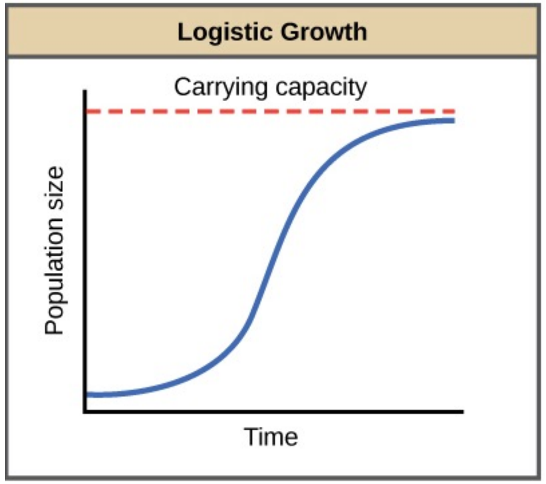
Modeling natural and phylogenetic phenomena: Bacterial Grow



Population size = 2^t



Population size = 2^t

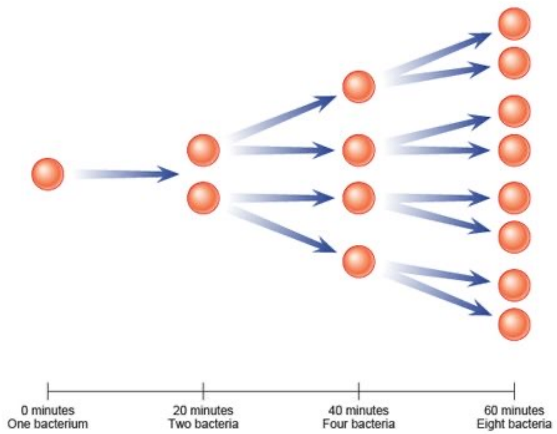


$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}}$$

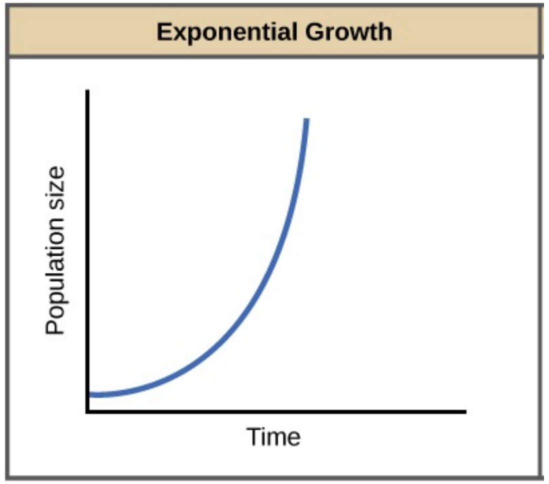
Models are sets of rules describing how a system changes over time

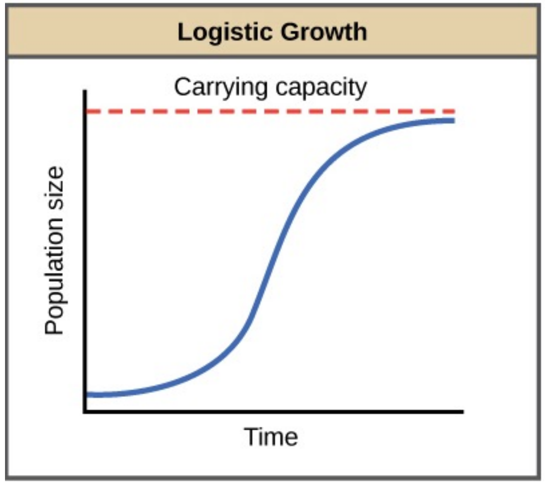
Modeling natural and phylogenetic phenomena: Bacterial Grow

Q: Do you know any other models used in biology?



Population size = 2^t

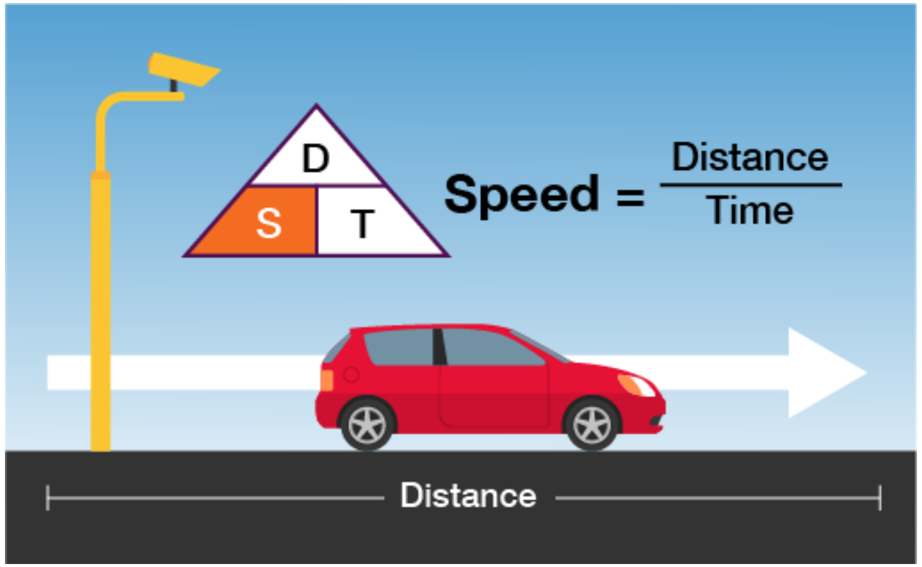




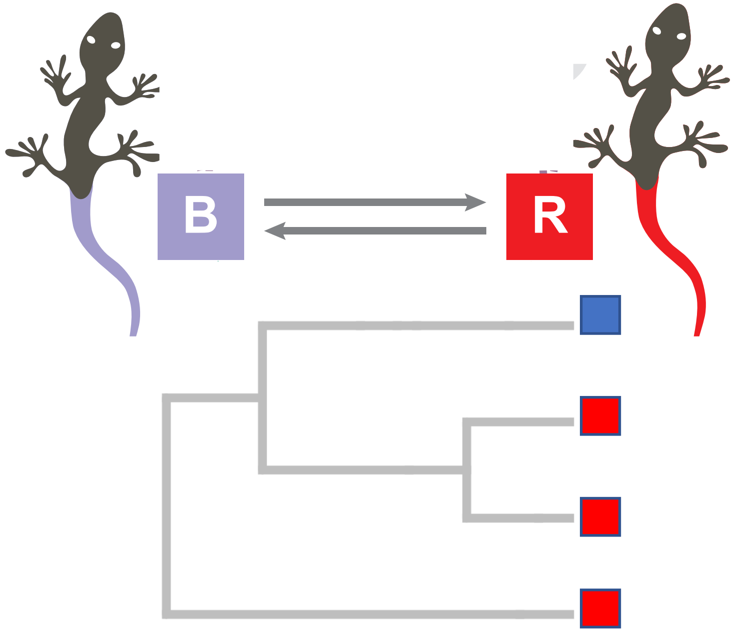
Population size = 2^t

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Modeling natural and phylogenetic phenomena



Models are common in physics



Models in biology?

Models are sets of rules describing how a system changes over time

Parsimony vs. Statistical Phylogenetics

Parsimony

Statistical models

Parsimony vs. Statistical Phylogenetics

Parsimony

- Strict principle of Occam’s razor
- Not a statistical model
- But can be considered as a mathematical model
- Some analysis are challenging (e.g., tree dating)
- FAST

Statistical models

Parsimony vs. Statistical Phylogenetics

Parsimony

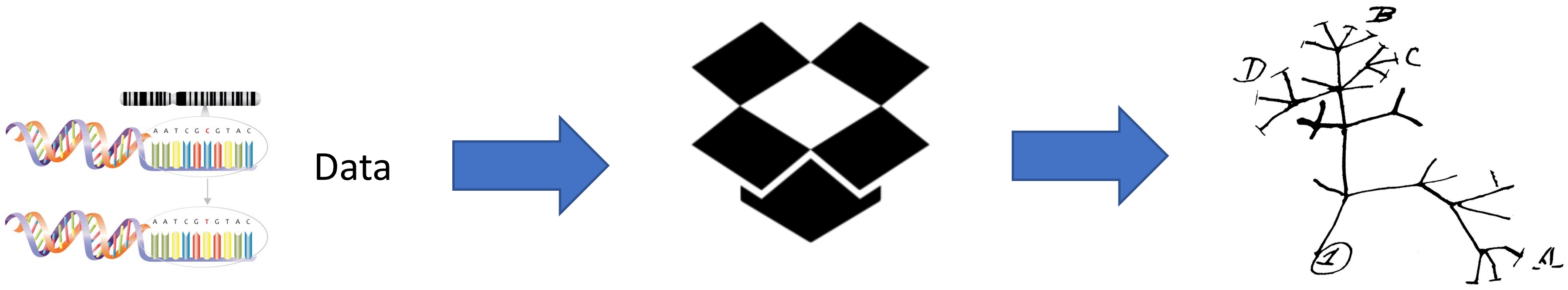
- Strict principle of Occam’s razor
- Not a statistical model
- But can be considered as a mathematical model
- Some analysis are challenging (e.g., tree dating)
- FAST

Statistical models

- Can model manifold of natural phenomena & processes
- Many different methods available
- Many different models to create
- Can be SLOW

Aim of this part of the course

- Main principles of modeling data

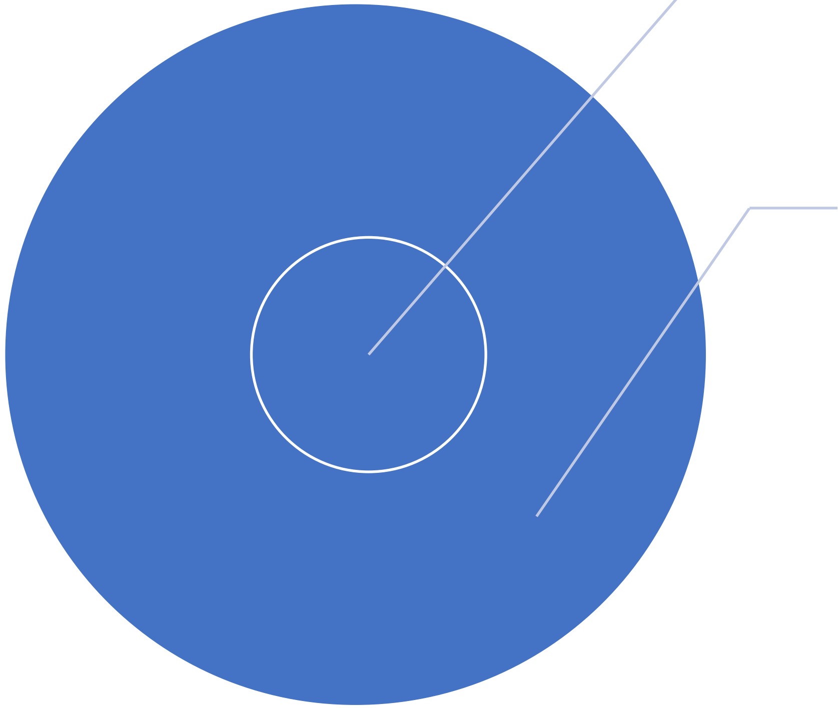


Data

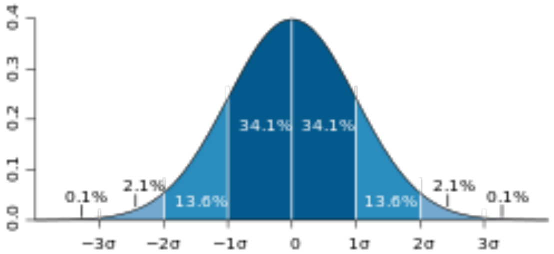
Inference

Modeling principles in phylogenetics

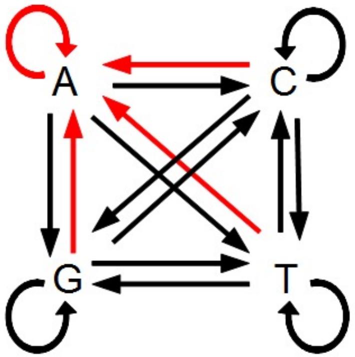
PHYLOGENETCIS



Probability Distribution



Stochastic models of character evolution (DNA or Morphology)



Intro to maximum likelihood (ML) method using simple Binomial model (coin toss)

- Rules of Probability
- Probability Distribution
- Binomial Model
- Likelihood of Binomial Model

Quick intro to probability: main rules

- AND rule

AND = 0.5*0.5 = 0.25

- OR rule

OR = 0.5 + 0.5 = 1

- Sum of all events

= 1

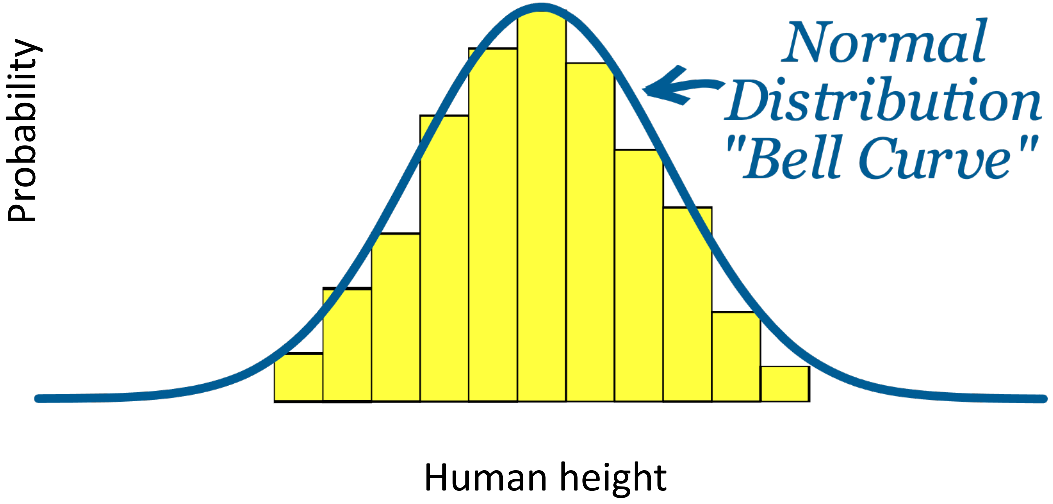
3 trials

Let’s play with a fair dice.

- Q1: Probability of seeing “1” after one trial?
  - 1/6
- Q2: Probability of seeing “2” AND “4” given two trials?
  - 1/6 * 1/6 = 1/36
- Q3: Probability of NOT seeing “3” given one trial?
  - 1– 1/6 = 5/6
- Q2: Probability of seeing “1” OR “6” given one trial?
  - 1/6 + 1/6 = 1/3

Probability distribution

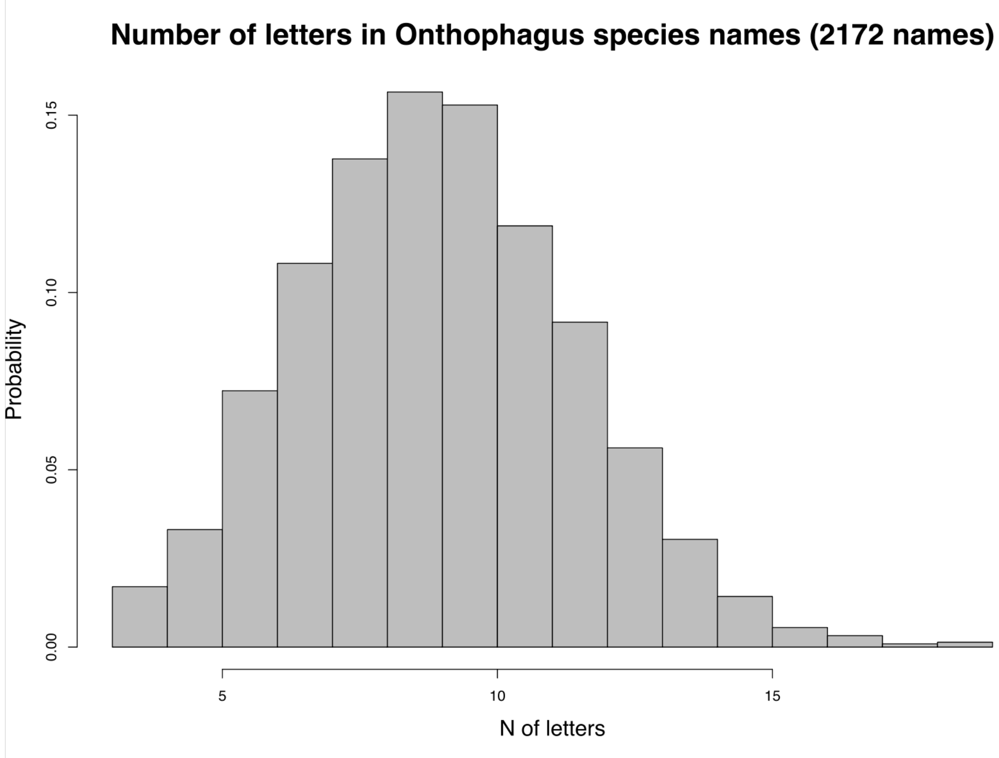
- A probability distribution is a function that provides the probabilities of occurrence of different possible outcomes in an experiment.
- Distribution usually refers to a distribution of a random variable
- In probability and statistics, a random variable, is a variable whose possible values are outcomes of a random phenomenon.



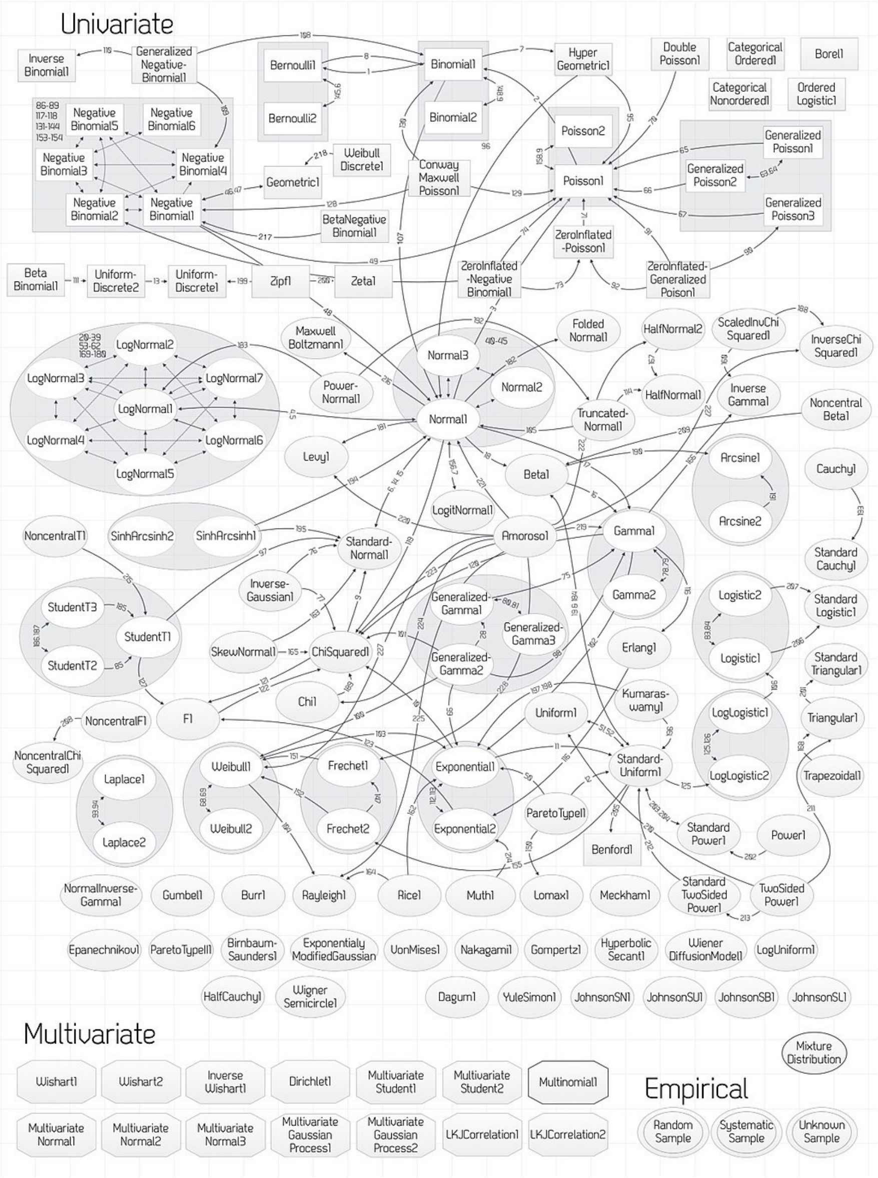
Informal Axioms (rules) of Statistics

- Any measured quantity of any set of objects in the Universe has some probability distribution
- There are ~20 most common distributions in the Universe (e.g., Binomial, Normal, Gamma, Poisson etc.)
- Most likely, the measured quantity falls into one of those ~20 common distributions

Empirical probability distributions



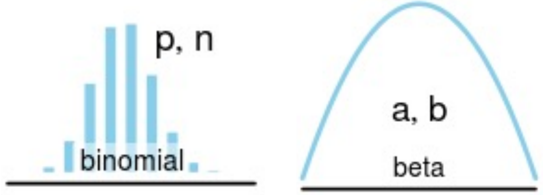
Relationships among probability distributions



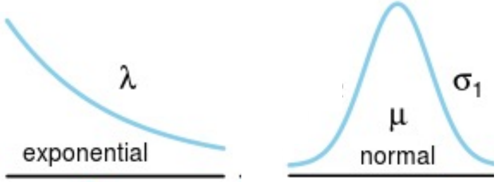
https://en.wikipedia.org/wiki/Relationships_among_probability_distributions

Classification of probability distributions

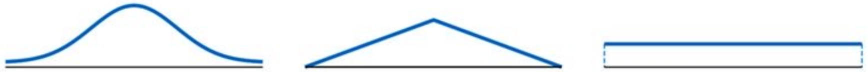
- Discrete vs. Continuous



- By number of parameters

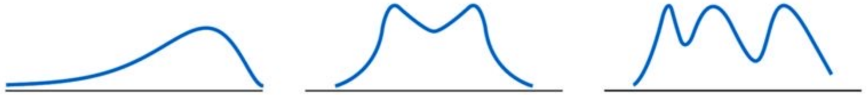


- By domain [-∞, +∞) vs. [0, +∞) vs. [0,1]
- By shape

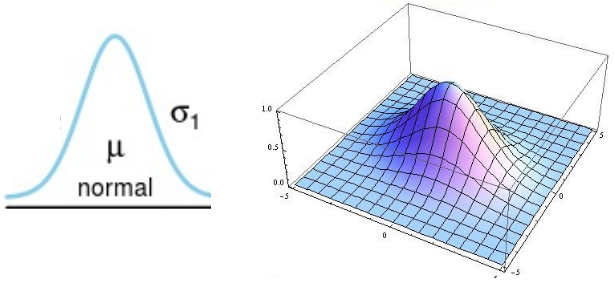


(a) Bell-shaped (b) Triangular (c) Uniform (or rectangular)

- By mode: unimodal vs. multimodal



- By dimension of random variable: univariate vs. multivariate



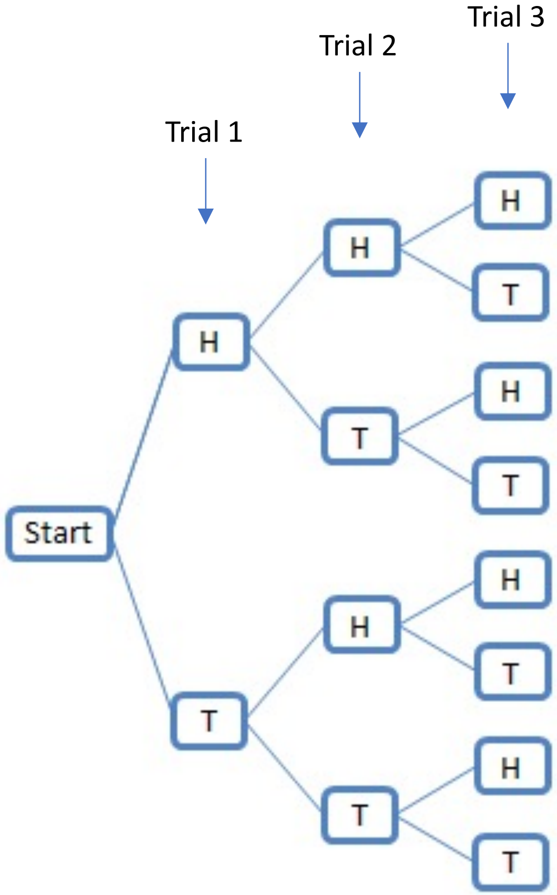
Binomial model (and distribution)

Binomial model gives the probability of seeing *k* heads in *n* coin tosses (trials) given that probability of seeing a head in one coin toss is *p*.

Let’s consider an example where *n*=3 and *p*=0.5

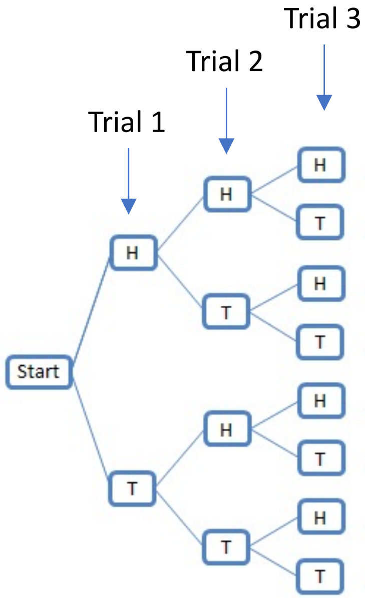
- Coin is fair
- We toss the coin 3 times

Estimating number of heads



Heads	Tails
3	0
2	1
2	1
1	2
2	1
1	2
1	2
0	3

Estimating probabilities



Heads	Tails
3	0
2	1
2	1
1	2
2	1
1	2
1	2
0	3

$= 0.5^3 = 0.125$

$= 0.5^3 = 0.125$

....

$= 0.5^3 = 0.125$

Estimating probabilities

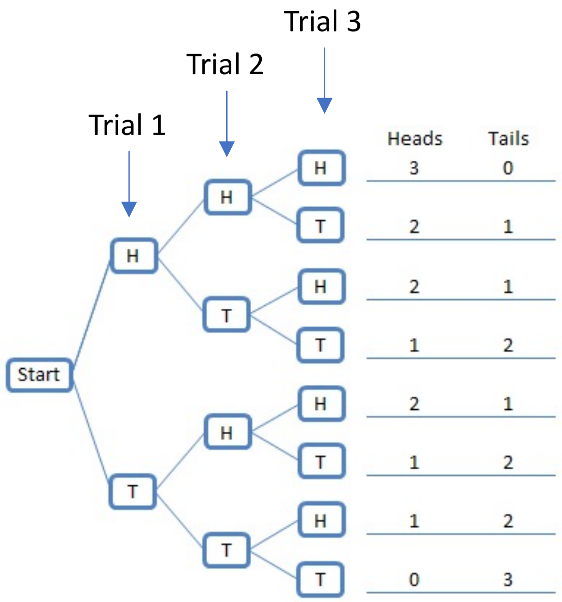
$$\boxed{\binom{n}{k} = \frac{n!}{k!\,(n-k)!}}$$

$$\binom{n=3}{k=3} = \frac{3!}{3!\,(3-3)!} = 1$$

$$\binom{n=3}{k=2} = \frac{3!}{2!\,(3-2)!} = 3$$

$$\binom{n=3}{k=1} = 3$$

$$\binom{n=3}{k=0} = 1$$



Heads	Tails
3	0
2	1
2	1
1	2
2	1
1	2
1	2
0	3

$= 0.5^3 = 0.125$

$= 0.5^3 = 0.125$

....

$= 0.5^3 = 0.125$

Estimating probabilities

Number *k* heads in 3 trials

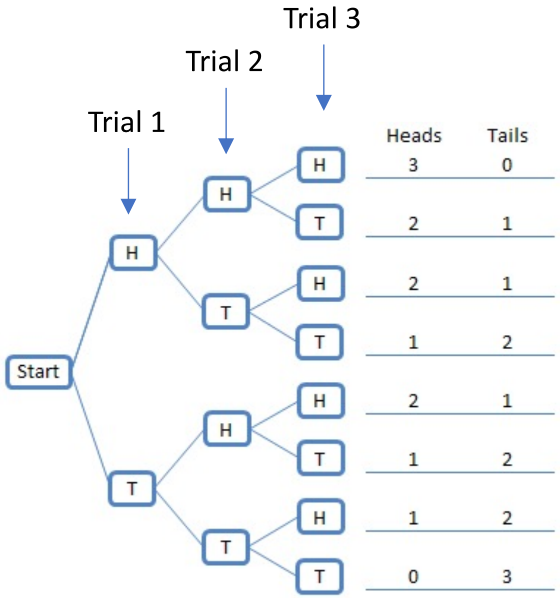
Probability of seeing *k* heads in 3 trials

$$\binom{n=3}{k=\mathbf{3}} = 1 \qquad P(3) = 1 * 0.5^3$$

$$\binom{n=3}{k=\mathbf{2}} = 3 \qquad P(2) = 3 * 0.5^3$$

$$\binom{n=3}{k=\mathbf{1}} = 3 \qquad P(1) = 3 * 0.5^3$$

$$\binom{n=3}{k=\mathbf{0}} = 1 \qquad P(0) = 1 * 0.5^3$$



Heads	Tails
3	0
2	1
2	1
1	2
2	1
1	2
1	2
0	3

$= 0.5^3 = 0.125$

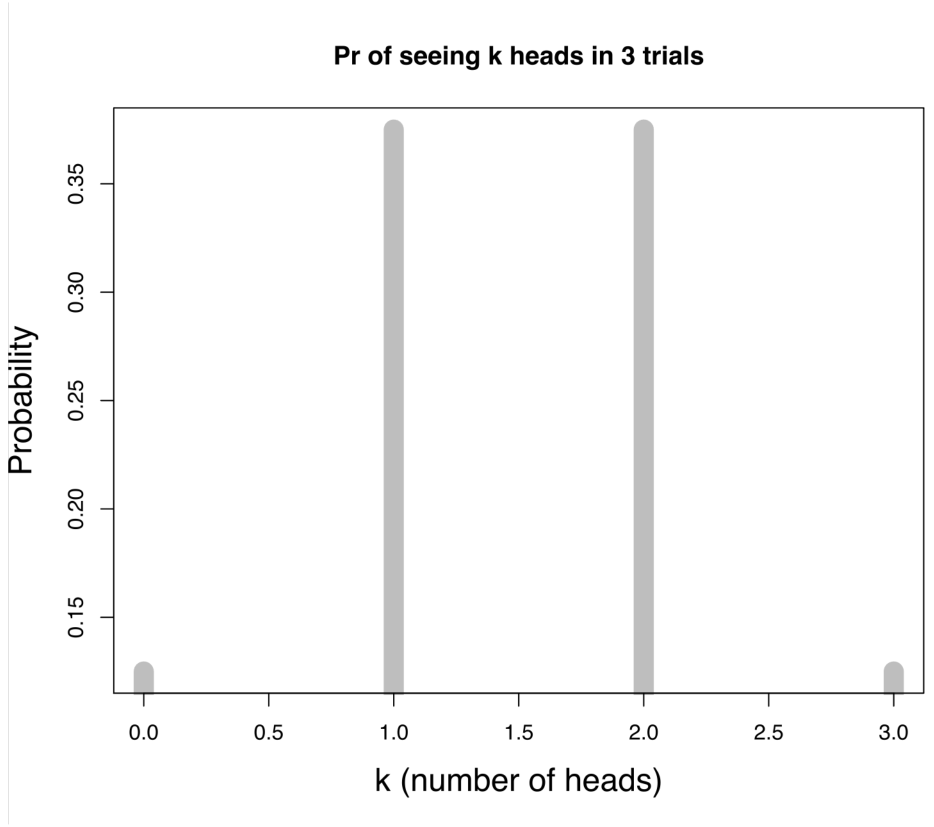
$= 0.5^3 = 0.125$

....

$= 0.5^3 = 0.125$

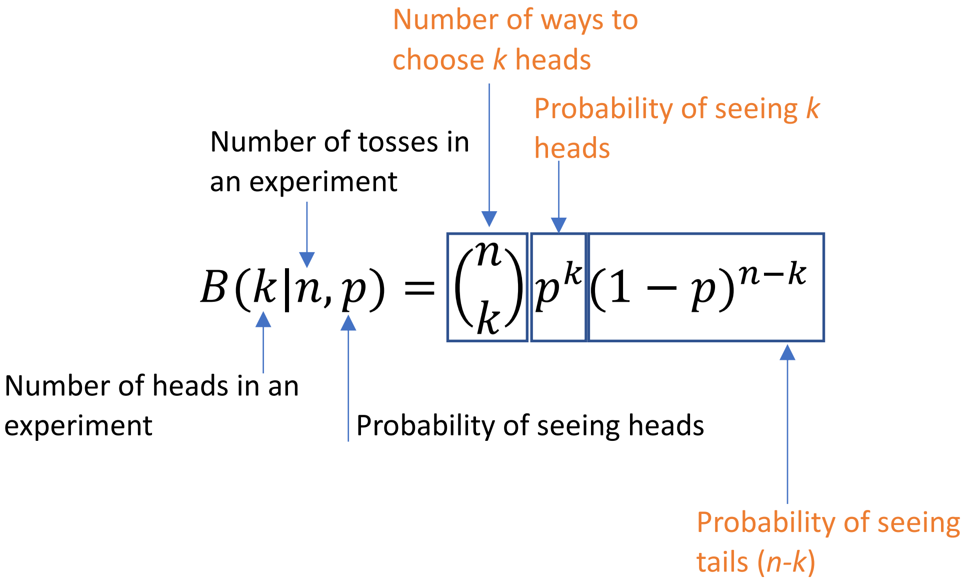
Estimating probabilities

Number *k* heads in 3 trials	Probability of seeing *k* heads in 3 trials
$\binom{n\,=\,3}{k\,=\,\mathbf{3}} = 1$	$P(3) = 1 * 0.5^3$
$\binom{n\,=\,3}{k\,=\,\mathbf{2}} = 3$	$P(2) = 3 * 0.5^3$
$\binom{n\,=\,3}{k\,=\,\mathbf{1}} = 3$	$P(1) = 3 * 0.5^3$
$\binom{n\,=\,3}{k\,=\,\mathbf{0}} = 1$	$P(0) = 1 * 0.5^3$



Binomial model (and distribution)

Binomial model gives the probability of seeing *k* heads in *n* coin tosses (trials) given that probability of seeing a head in one coin toss is *p*.



Number of ways to choose *k* heads

Probability of seeing *k* heads

Number of tosses in an experiment

$$B(k|n,p) = \boxed{\binom{n}{k}}\boxed{p^k}\boxed{(1-p)^{n-k}}$$

Number of heads in an experiment

Probability of seeing heads

Probability of seeing tails (*n*-*k*)

$$\boxed{\binom{n}{k} = \frac{n!}{k!\,(n-k)!}}$$

- Coin is fair
- We toss the coin 3 times

Binomial model

Number of ways to choose 2 heads in 3 tosses = 3

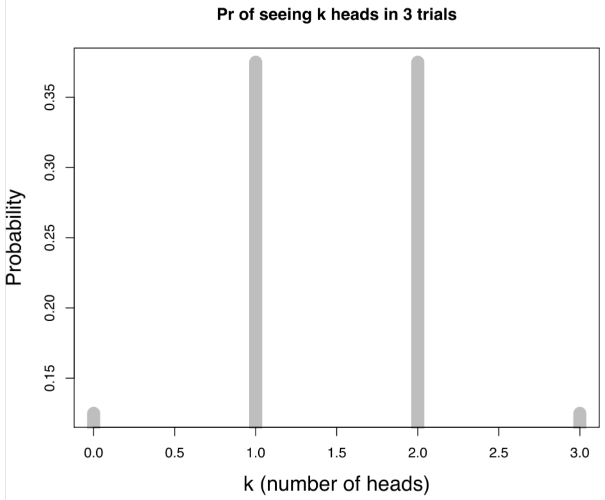
Number of tosses in an experiment

Probability of seeing 2 heads in 3 trials

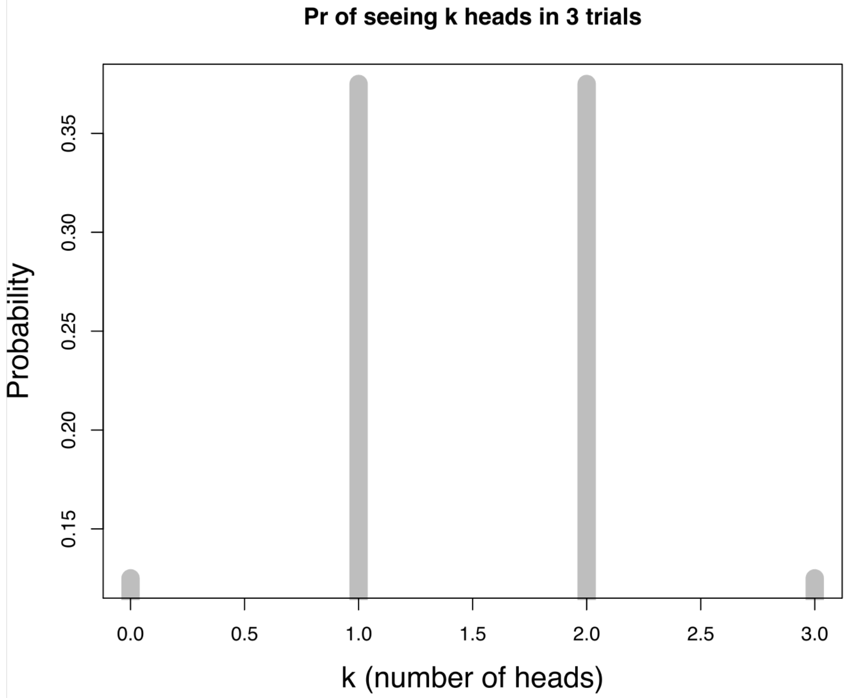
$$B(2\,|3,0.5) = \boxed{\binom{3}{2}}\underbrace{0.5^2(1-0.5)^{3-2}}_{\text{Probability of seeing 2 heads and 1 tail} = 0.5^2 * 0.5} = 3 * 0.5^2 * 0.5 \quad = 0.375$$

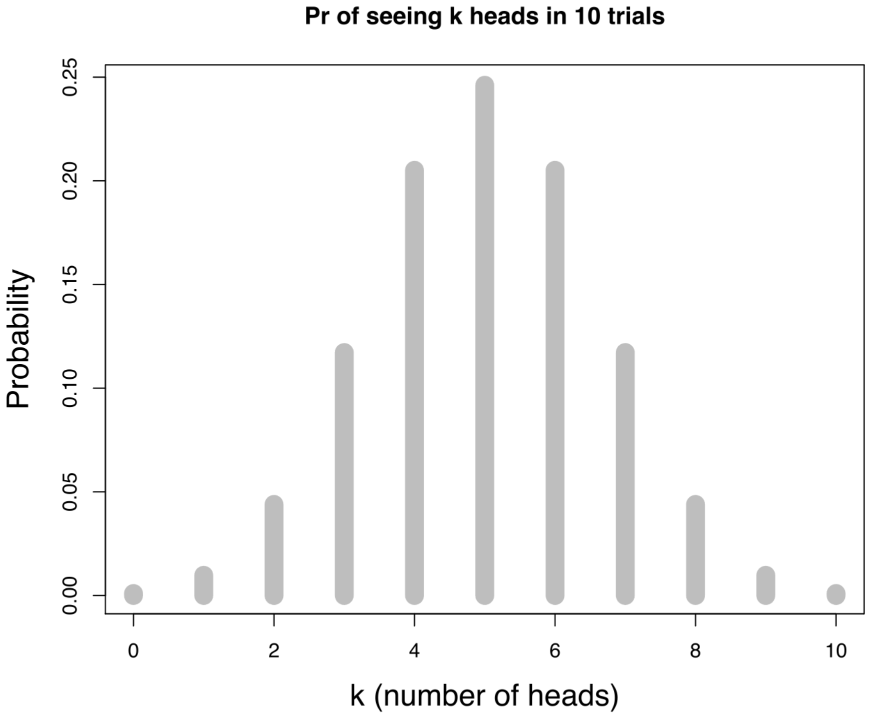
Number of heads in an experiment

Probability of seeing heads



Binomial model





Likelihood function of Binomial distribution

In statistics, a likelihood function (often simply the likelihood) is a particular function of the parameter of a statistical model given data. Likelihood functions play a key role in statistical inference.

Binomial model:
P(k) of *k* heads given *n* trials and *p* of seeing a head in a trial

$$B(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

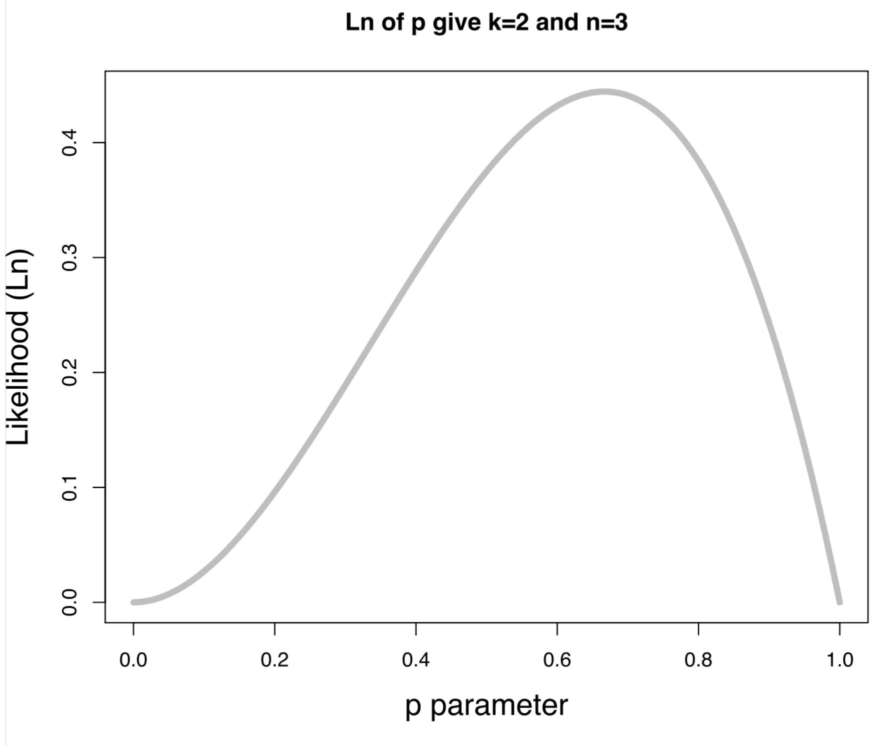
Binomial Likelihood:
Given *n* and *k* infer *p* that maximizes the likelihood function

$$Ln(p \mid n = 3, k = 2) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{3}{2} p^2 (1-p)^{3-2}$$

- Domain of *p* is a value between 0 and 1 (since *p* is a probability)
- Let’s try all *p*’s to get a likelihood function
- Likelihood function is not a distribution

Likelihood function of Binomial distribution

In statistics, a likelihood function (often simply the likelihood) is a particular function of the parameter of a statistical model given data. Likelihood functions play a key role in statistical inference.



Binomial Likelihood:
Given *n* and *k* infer *p* that maximizes the likelihood function

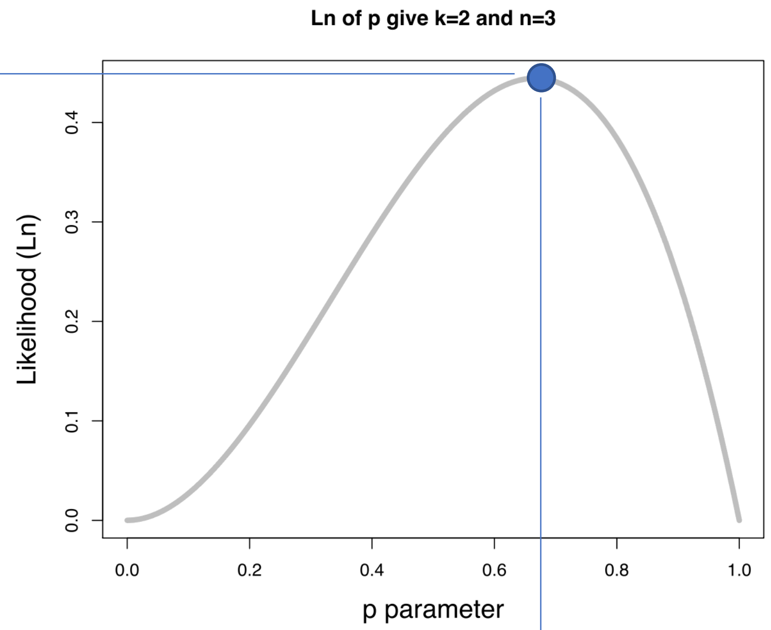
$$Ln(p \mid n = 3, k = 2) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{3}{2} p^2 (1-p)^{3-2}$$

- Domain of *p* is a value between 0 and 1 (since *p* is a probability)
- Let's try all *p*'s to get a likelihood function
- Likelihood function is not a distribution

Maximum Likelihood

Maximum value of the likelihood function

Ln=0.44

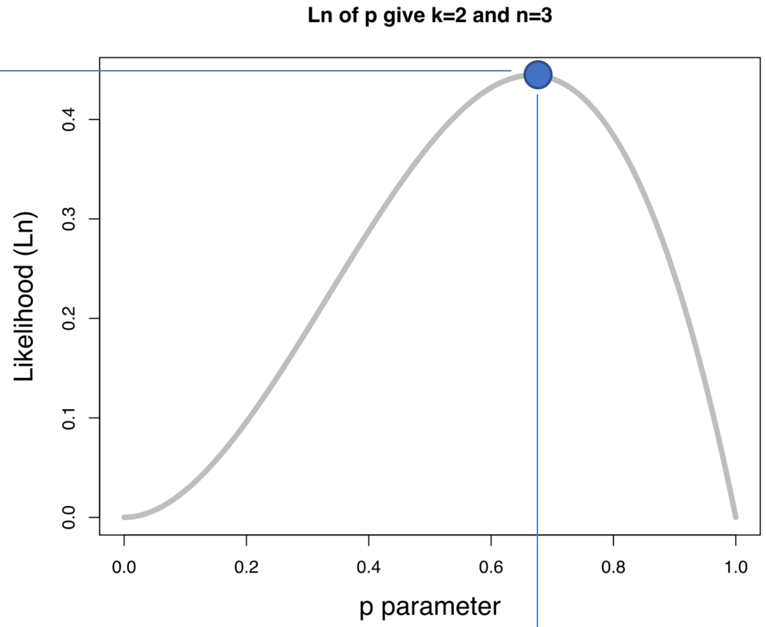


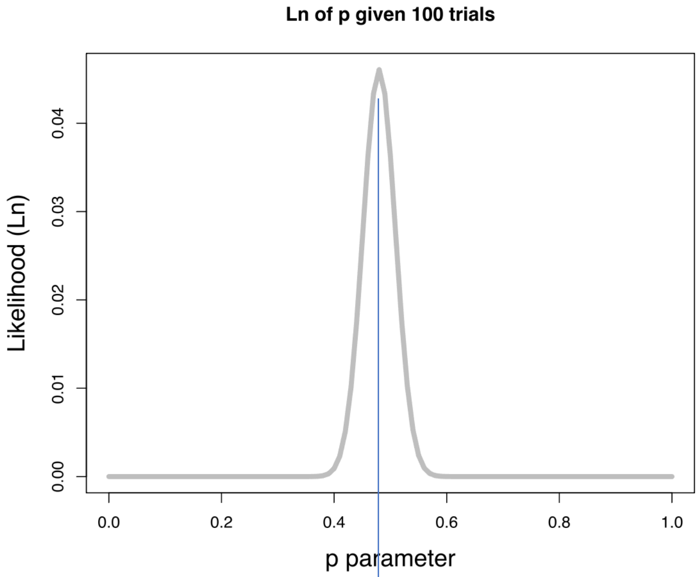
Maximum likelihood estimate of the parameter

$Ln(\hat{p}) = \frac{k}{n} = \frac{2}{3}$ = 0.667

Maximum Likelihood

Maximum value of the likelihood function *Ln*=0.44





Maximum likelihood estimate of the parameter

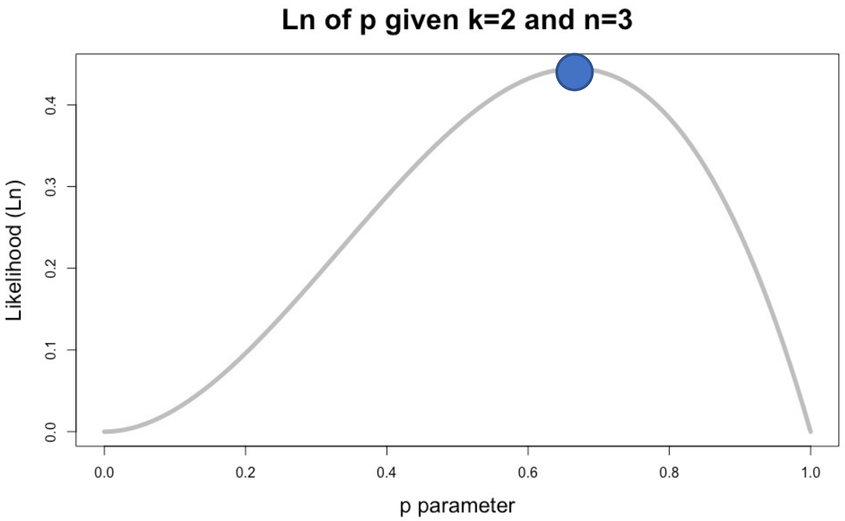
$Ln(\hat{p}) = \frac{k}{n} = \frac{2}{3} =$ 0.667

$Ln(\hat{p})$ =0. 48

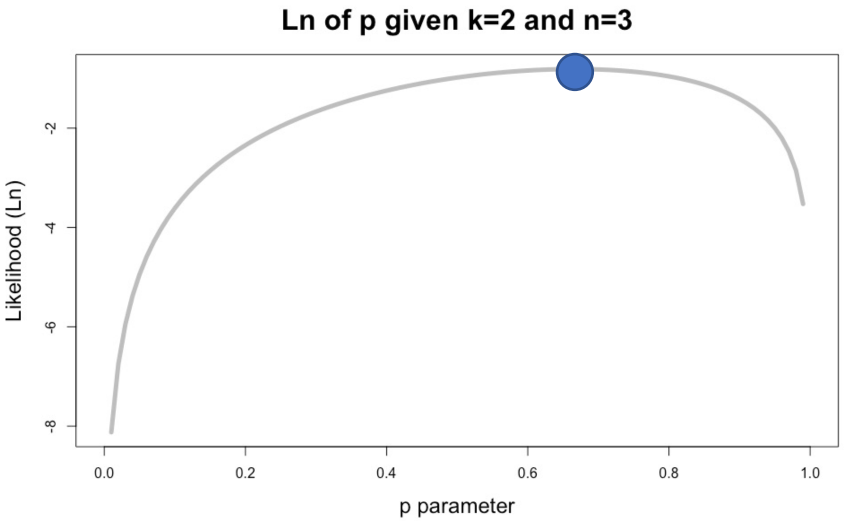
Likelihood properties

- Likelihood is not a probability distribution!
- Log likelihood

likelihood



Log likelihood

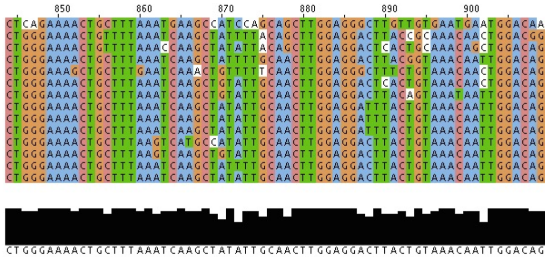


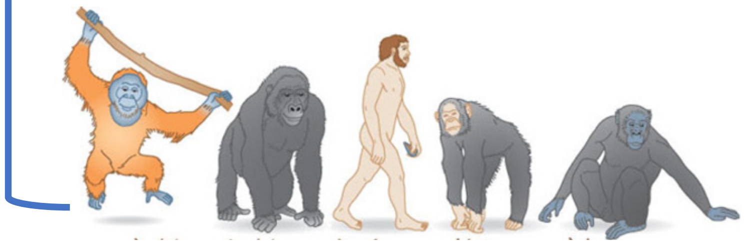
Summary

- Statistical phylogenetics is about modeling evolutionary process using probability distribution and stochastic processes
- Every measurement in this world is roughly speaking is a realization of some stochastic process
- In other words: every measurement is an instance that comes from some probability distribution (= model)
- Models are set of rules that describe how systems evolve
- In modeling data we need to come up with models that realistically describe our world
- **Maximum Likelihood method:** given that we observe an outcome and know the generating model, we can estimate the parameters of the process.

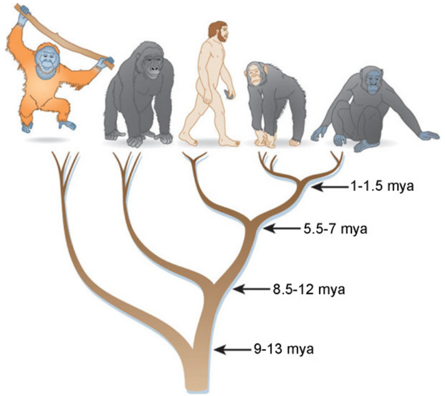
Tomorrow’s lecture

Observed data (DNA & Morphology)





inference



Let’s calculate likelihood of some coin?