First steps in *BEAST*2 and setting priors

Chi Zhang

Installation

Beast2
Bayesian evolutionary analysis by sampling trees

https://www.beast2.org

FigTree is a program for viewing trees, publication quality figures.

https://beast.community/figtree

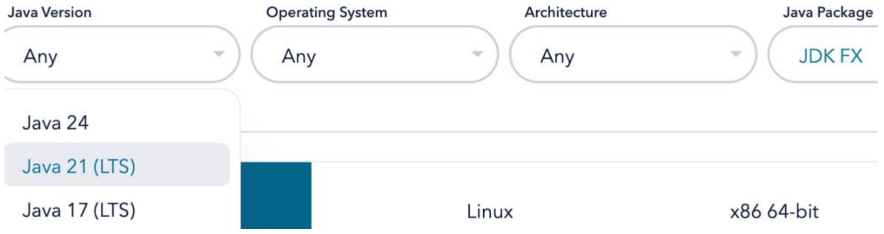
Tracer is a graphical tool for visualization and diagnostics

https://beast.community/tracer

Looking just for Java?

Azul Zulu Builds of OpenJDK

https://www.azul.com/downloads/?package=jdk-fx



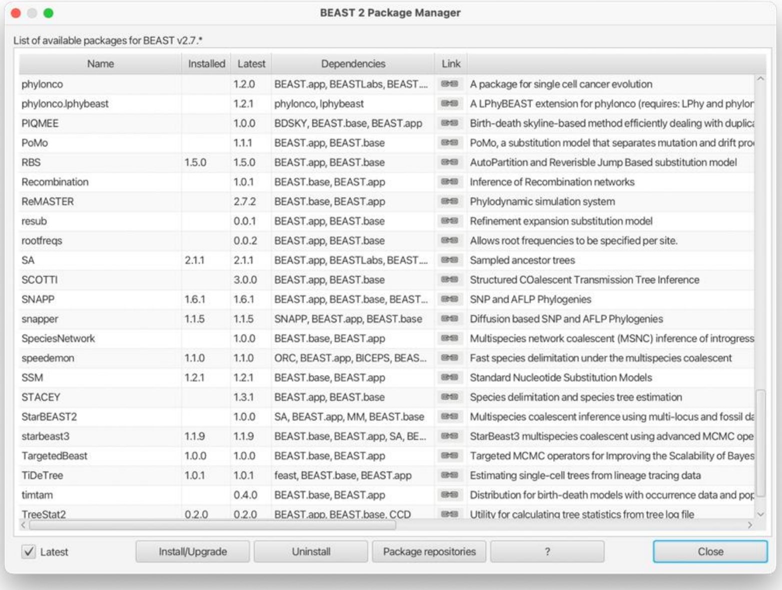
The R Project for Statistical Computing

https://www.r-project.org
https://posit.co/download/rstudio-desktop

Installation

- Managing packages

  https://www.beast2.org/managing-packages/index.html



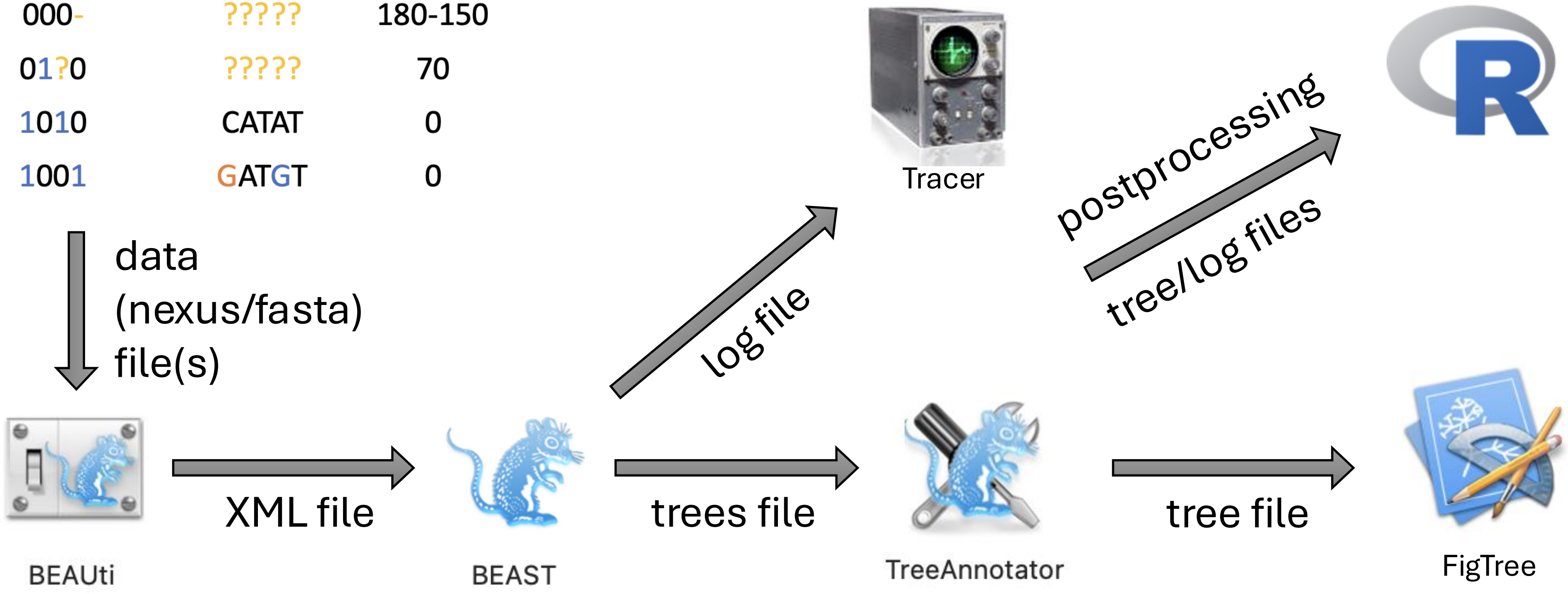
Name	Installed	Latest	Dependencies	Link	
phylonco		1.2.0	BEAST.app, BEASTLabs, BEAST....		A package for single cell cancer evolution
phylonco.lphybeast		1.2.1	phylonco, lphybeast		A LPhyBEAST extension for phylonco (requires: LPhy and phylor
PIQMEE		1.0.0	BDSKY, BEAST.base, BEAST.app		Birth-death skyline-based method efficiently dealing with duplic
PoMo		1.1.1	BEAST.app, BEAST.base		PoMo, a substitution model that separates mutation and drift pro
RBS	1.5.0	1.5.0	BEAST.app, BEAST.base		AutoPartition and Reverisble Jump Based substitution model
Recombination		1.0.1	BEAST.base, BEAST.app		Inference of Recombination networks
ReMASTER		2.7.2	BEAST.base, BEAST.app		Phylodynamic simulation system
resub		0.0.1	BEAST.app, BEAST.base		Refinement expansion substitution model
rootfreqs		0.0.2	BEAST.app, BEAST.base		Allows root frequencies to be specified per site.
SA	2.1.1	2.1.1	BEAST.app, BEASTLabs, BEAST....		Sampled ancestor trees
SCOTTI		3.0.0	BEAST.app, BEAST.base		Structured COalescent Transmission Tree Inference
SNAPP	1.6.1	1.6.1	BEAST.app, BEAST.base, BEAST...		SNP and AFLP Phylogenies
snapper	1.1.5	1.1.5	SNAPP, BEAST.app, BEAST.base		Diffusion based SNP and AFLP Phylogenies
SpeciesNetwork		1.0.0	BEAST.base, BEAST.app		Multispecies network coalescent (MSNC) inference of introgress
speedemon	1.1.0	1.1.0	ORC, BEAST.app, BICEPS, BEAS...		Fast species delimitation under the multispecies coalescent
SSM	1.2.1	1.2.1	BEAST.base, BEAST.app		Standard Nucleotide Substitution Models
STACEY		1.3.1	BEAST.app, BEAST.base		Species delimitation and species tree estimation
StarBEAST2		1.0.0	SA, BEAST.app, MM, BEAST.base		Multispecies coalescent inference using multi-locus and fossil da
starbeast3	1.1.9	1.1.9	BEAST.base, BEAST.app, SA, BE...		StarBeast3 multispecies coalescent using advanced MCMC ope
TargetedBeast	1.0.0	1.0.0	BEAST.base, BEAST.app		Targeted MCMC operators for Improving the Scalability of Bayes
TiDeTree	1.0.1	1.0.1	feast, BEAST.base, BEAST.app		Estimating single-cell trees from lineage tracing data
timtam		0.4.0	BEAST.base, BEAST.app		Distribution for birth-death models with occurrence data and pop
TreeStat2	0.2.0	0.2.0	BEAST.app, BEAST.base, CCD		Utility for calculating tree statistics from tree log file

- Adding package repositories

- Installing packages by hand
  - Windows: Users\<YourName>\BEAST\2.7\<PKG>
  - Mac: /Users/<YourName>/Library/Application Support/BEAST/2.7/<PKG>
  - Linux: /home/<YourName>/.beast/2.7/<PKG>

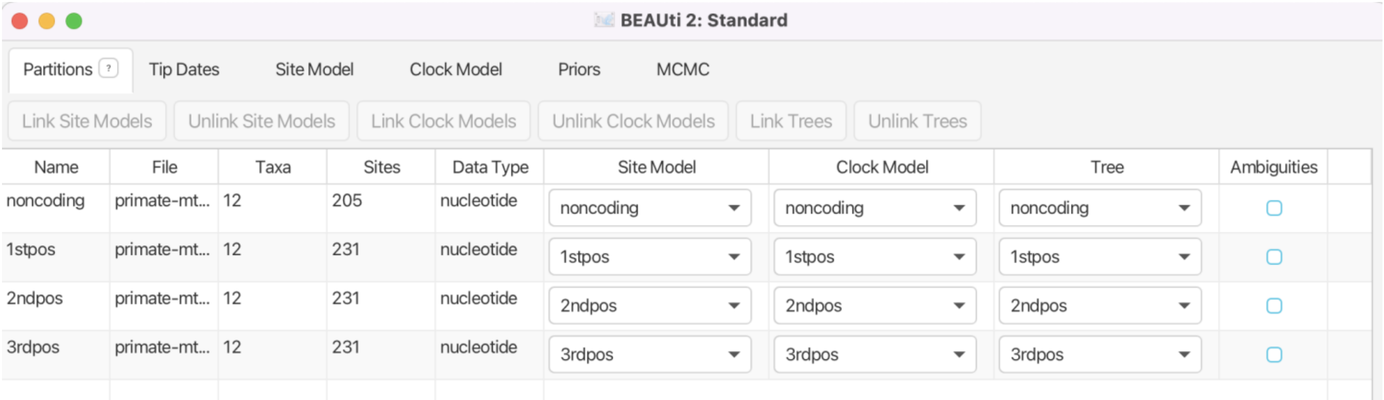
Workflow

Taxa	Morphological	Molecular	Age(Ma)
F_1	000-	?????	180-150
F_2	01?0	?????	70
S_1	1010	CATAT	0
S_2	1001	GATGT	0



Workflow

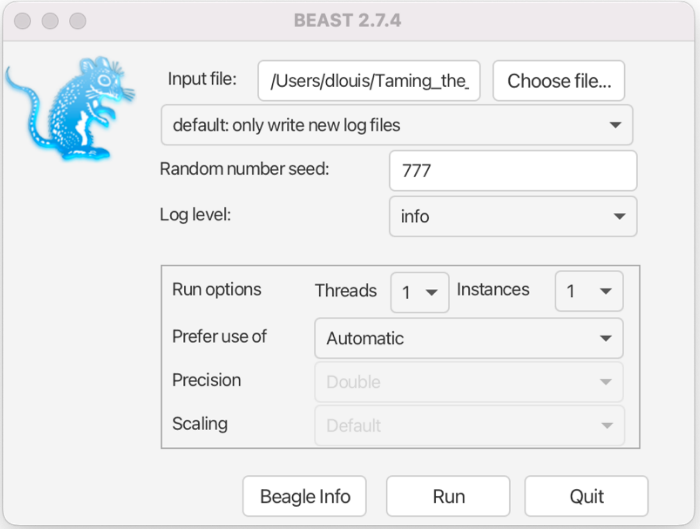
- https://taming-the-beast.org/tutorials/Introduction-to-BEAST2
- primate-mtDNA.nex
  - non-coding region
  - 1st codon positions
  - 2nd codon positions
  - 3rd codon positions



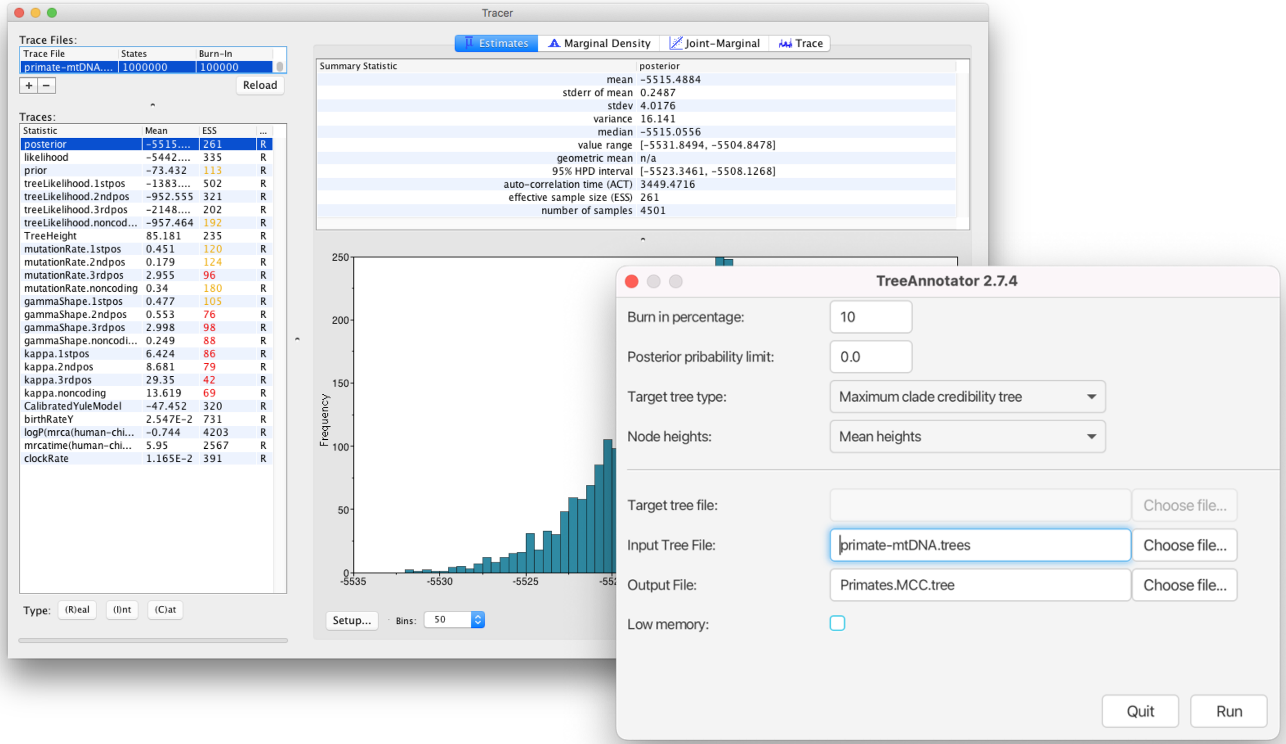
- Setting up the analysis
- Generating a XML file

Workflow

- Running the analysis

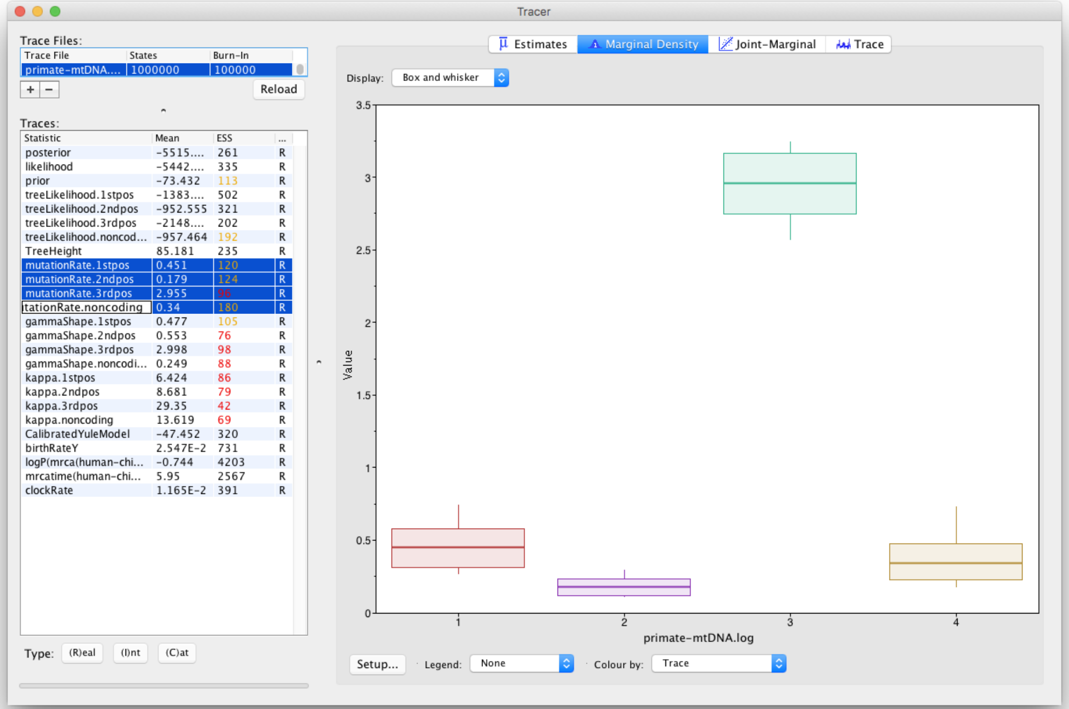


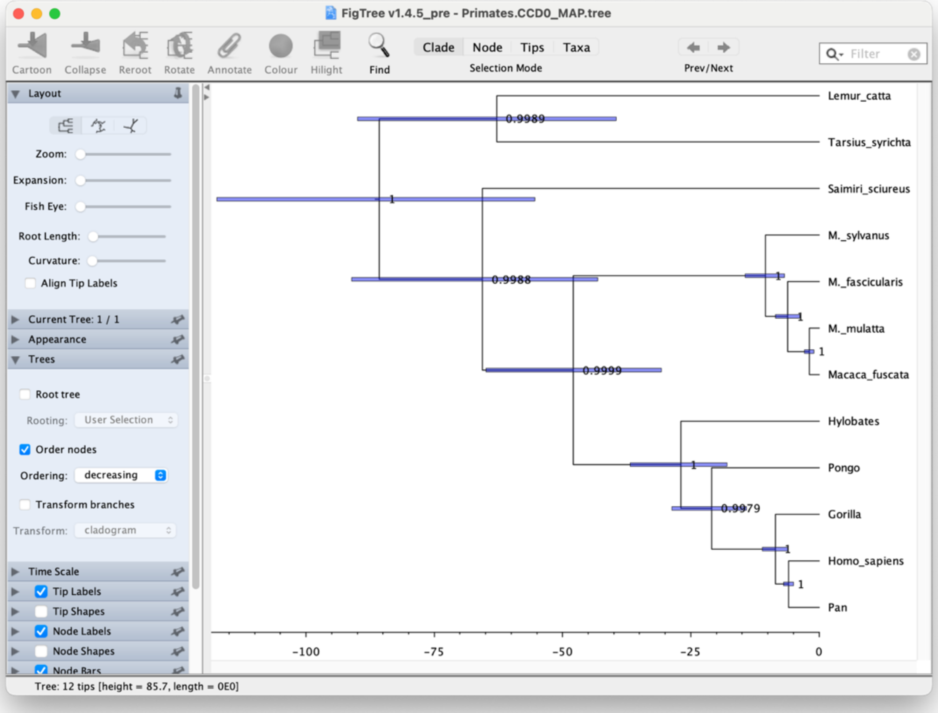
- Processing the results



Workflow

- Visualizing the results





Bayesian inference

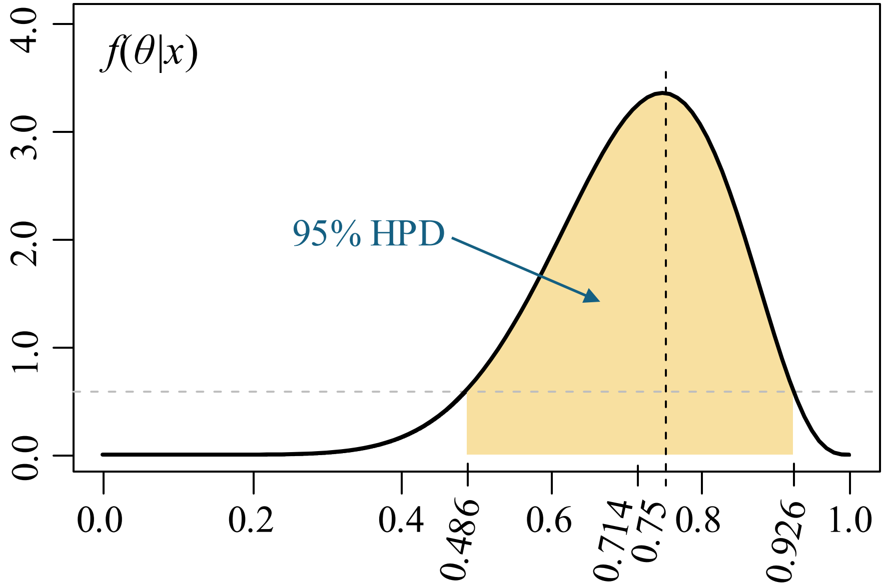
- Example: coin tossing (Yang 2014. p34)
- Data: x = 9 heads and r = 3 tails in n = 12 independent tosses
- Parameter: probability of head θ
- Likelihood: binomial distribution

$$\Pr(x|\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}$$

- Prior: $\theta \sim U(0, 1)$
- Posterior: $\theta|x \sim \mathrm{Beta}(10, 4)$

Bayesian inference

- Point estimates:
  - Mean: $10/14 = 0.714$
  - Median: $\approx (10 - 1/3) / (14 - 2/3) = 0.725$
- Credibility interval:
  - 95% equal-tail credibility interval (0.462, 0.909)
  - 95% highest posterior density interval (0.486, 0.926)



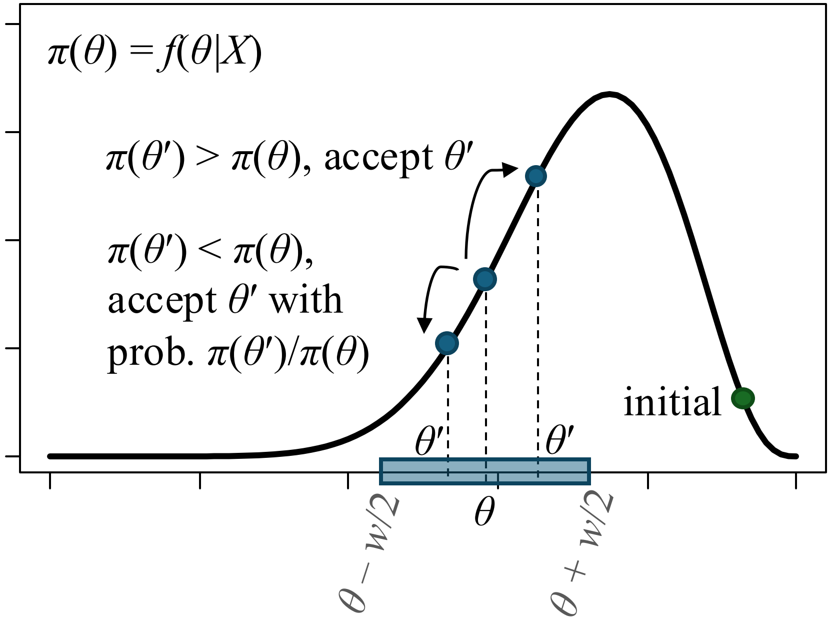
Prior

- Represent one's subjective belief about the parameter before seeing or analyzing the data (subjective Bayesian)
- Noninformative priors do not exist (uniform priors are not noninformative)
- Vague or diffuse priors
- Informative priors
- In practice: assess the influence of the prior

Markov chain Monte Carlo (MCMC)

- Metropolis algorithm (Metropolis et al. 1953)
  1. Set initial state of θ
  2. Propose a new state $\theta' \sim U(\theta - w/2, \theta + w/2)$
  3. If $\pi(\theta') > \pi(\theta)$, accept θ'; otherwise, accept θ' with probability $\alpha = \pi(\theta')/\pi(\theta)$
  4. If the proposal is accepted, set $\theta = \theta'$; otherwise, set $\theta = \theta$. Print out θ
  5. Go to step 2
- Acceptance ratio

$$\alpha = \min\left(1, \frac{\pi(\theta')}{\pi(\theta)}\right) = \min\left(1, \frac{f(\theta')f(X|\theta')}{f(\theta)f(X|\theta)}\right)$$



MCMC

- Metropolis-Hastings algorithm (Hastings 1970)
- Acceptance ratio $\alpha = \min\left(1, \frac{f(\theta')f(X|\theta')q(\theta|\theta')}{f(\theta)f(X|\theta)q(\theta'|\theta)}\right)$

prior ratio | likelihood ratio | proposal ratio

- Multiplier proposal
  - Propose a new state $\theta' = \theta c = \theta e^{\varepsilon(u - 1/2)}$
  - Proposal ratio is $c = \theta'/\theta$

MCMC

- Example: JC69 distance (Yang 2014. p216)
- Data: human and orangutan 12s rRNA genes (n = 948, x = 90)
- Likelihood: $f(x|d) = (p_1/4)^x (p_0/4)^{n-x}$

$$= \left(\frac{1}{16} - \frac{1}{16}\mathrm{e}^{-4d/3}\right)^x \left(\frac{1}{16} + \frac{3}{16}\mathrm{e}^{-4d/3}\right)^{n-x}$$

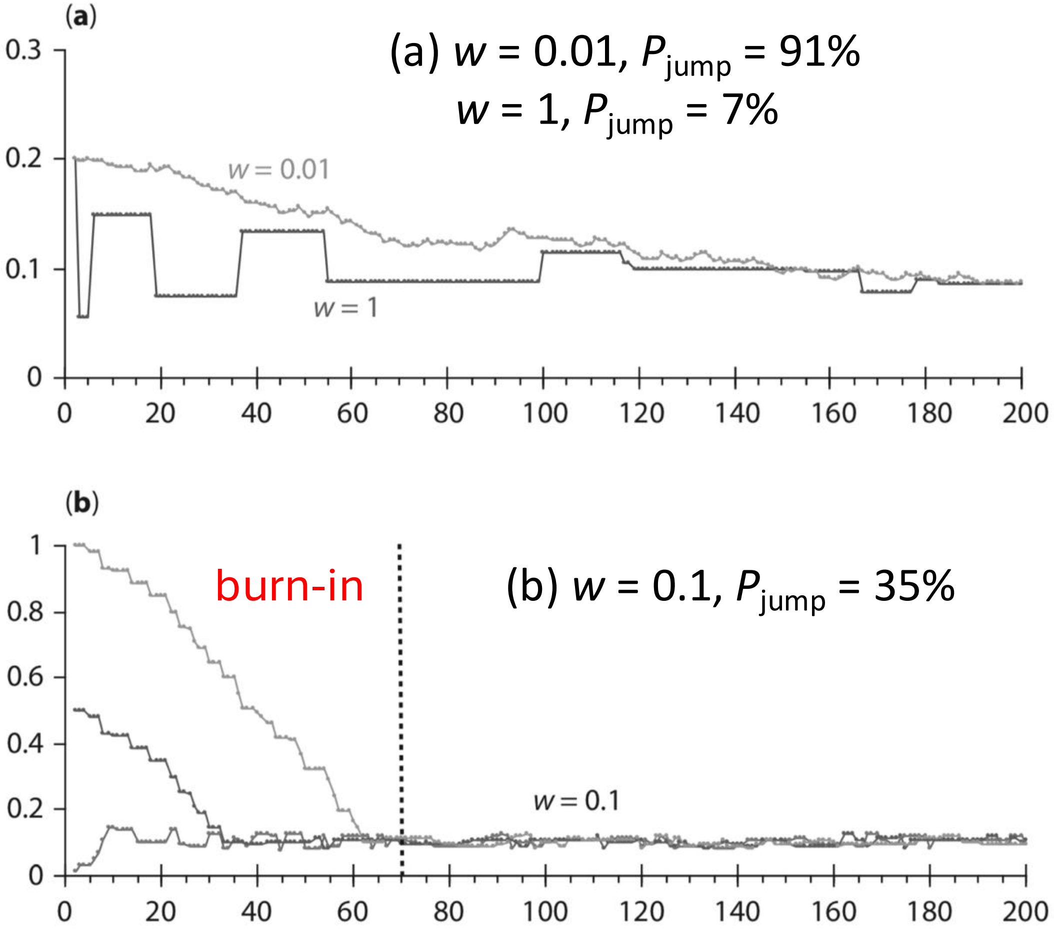
- Prior: $f(d) = \lambda \mathrm{e}^{-\lambda d}$, with $\lambda = 5$ (exponential distribution)

MCMC

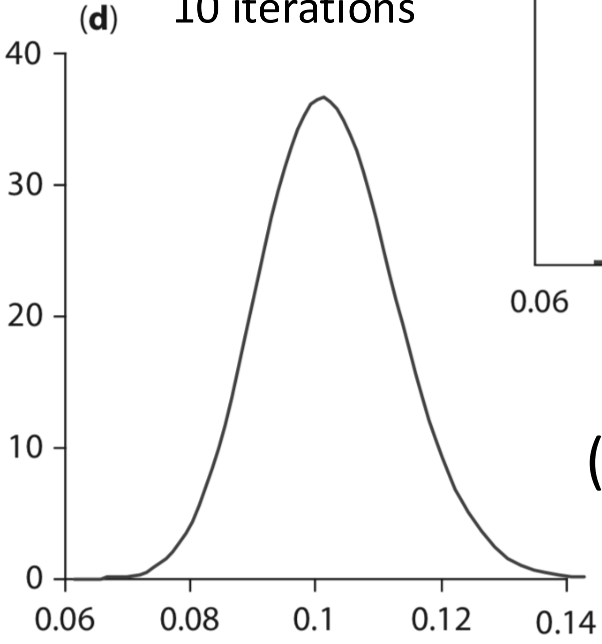
(a) w = 0.01, P_{jump} = 91%
w = 1, P_{jump} = 7%

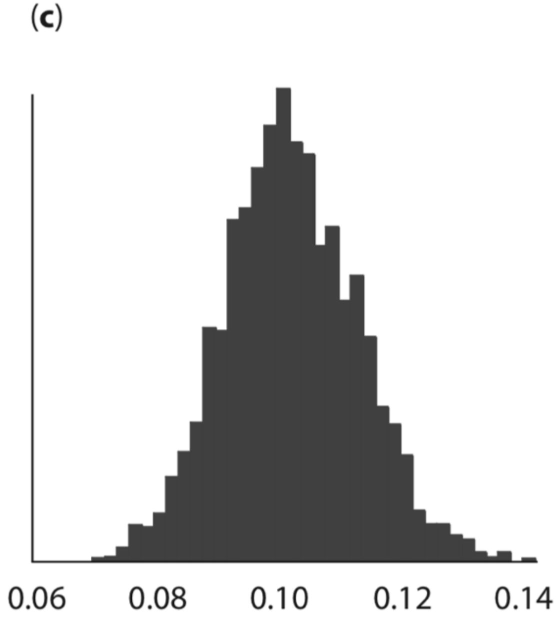
(b) w = 0.1, P_{jump} = 35%

burn-in



(c) 10^4 iterations
(d) 10^7 iterations
sampled every
10 iterations

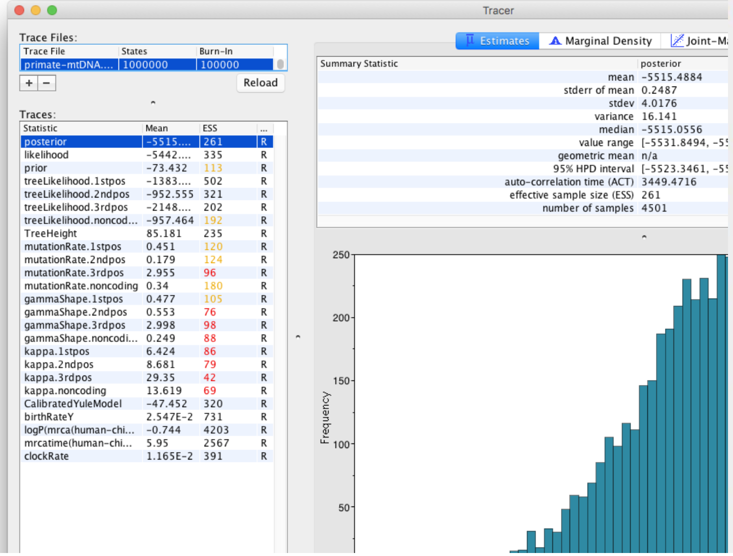


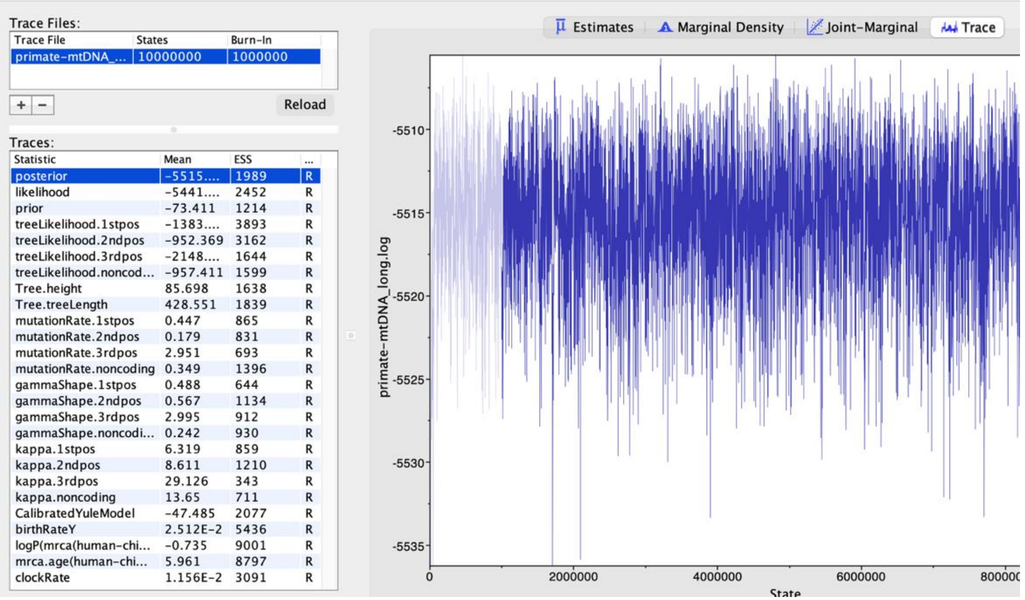


(Yang 2014. p217)

MCMC

- Effective sample size (ESS)
- A dependent sample of size N is as informative as an independent sample of size $N/[1+2(\rho_1+\rho_2+\rho_3+\cdots)]$.
- ESS > 200



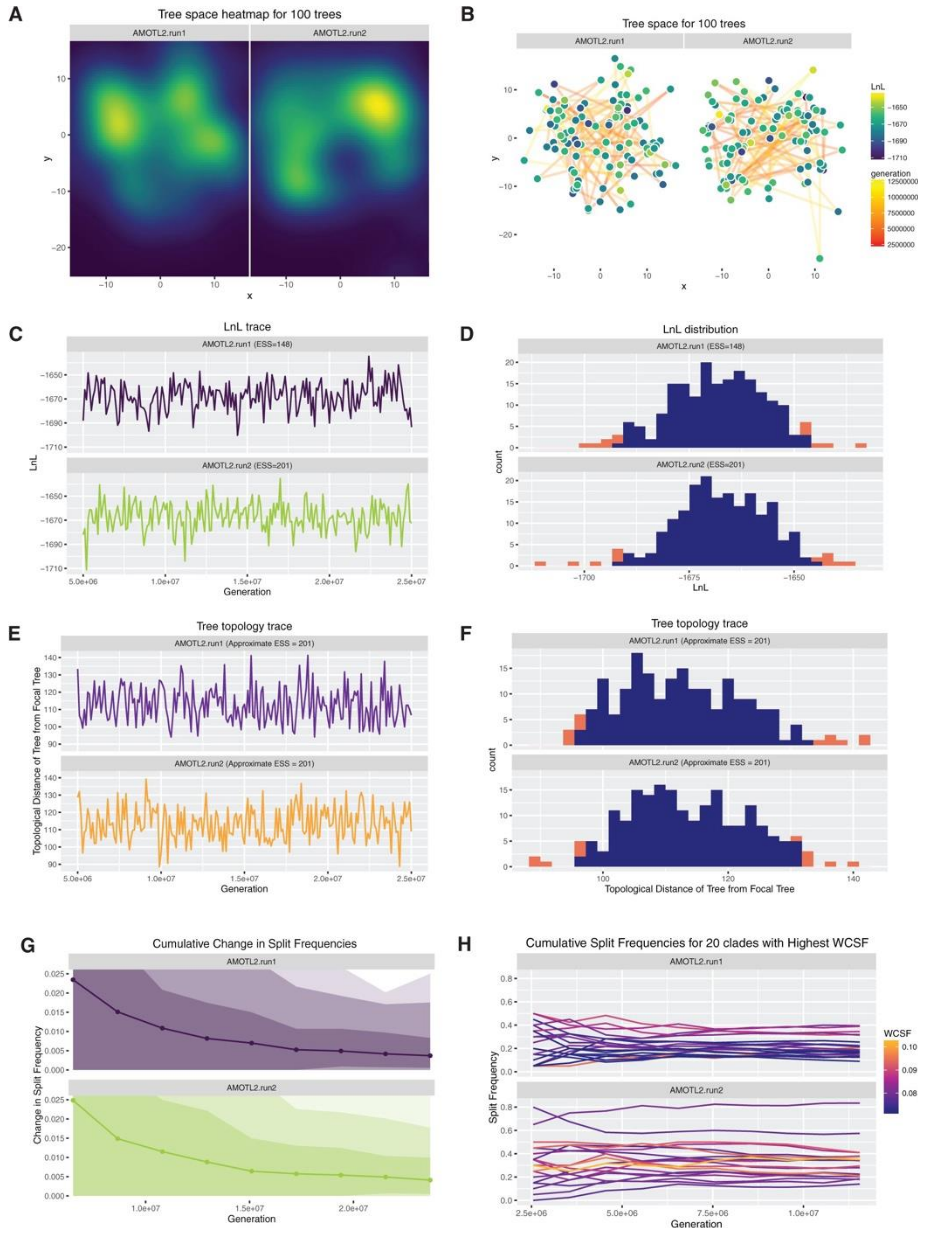


MCMC

- AWTY (R We There Yet) in R

  (Warren et al. 2017)

  - trace plots of parameters
  - (approx.) ESS of tree topologies
  - visualizing treespace
  - split frequencies



MCMC

- Run multiple independent chains
- Compare samples among runs
- Check convergence and mixing
- Combine the samples (logcombiner)

Summarizing trees

- Maximum *a posterior* (MAP) tree
  - tree topology with the maximum posterior probability
- Maximum clade credibility (MCC) tree
  - tree with the maximum product of clade probabilities
- Conditional clade distribution (CCD) MAP tree
  - clade frequencies (CCD0)
  - clade split frequencies (CCD1)
  - pairs of clade split frequencies (CCD2)
- CCD0-MAP tree should be the preferred point estimator (Berling et al. 2025) [does not work for trees with sampled ancestors]

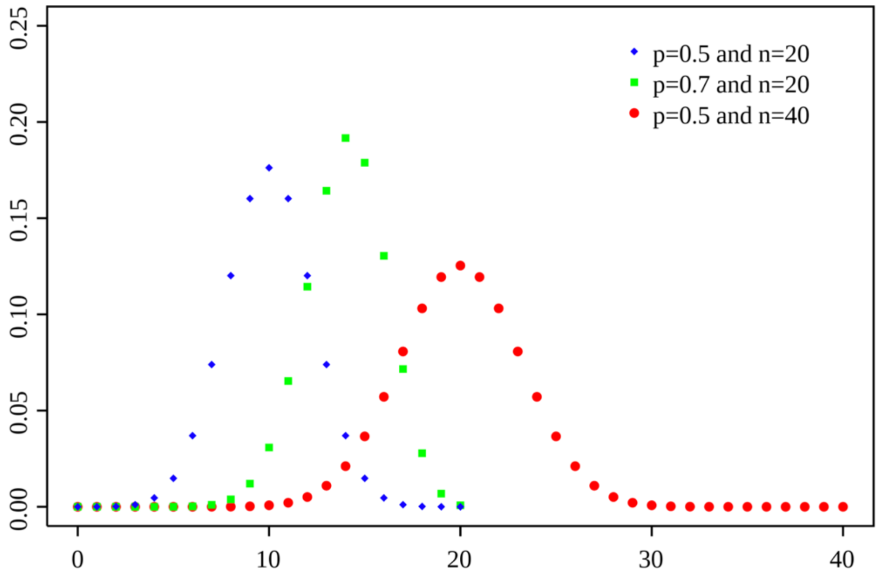
Probability distributions

- Discrete random variable
  - probability mass function
  - $\Pr(X = x_i) = p_i$ ($i = 1, 2, \ldots, k$, and k can be ∞), with $\sum_i p_i = 1$
- Continuous random variable
  - probability density function (PDF)
  - $\Pr(a < X < b) = \int_a^b f(x)dx$
  - area under the entire PDF curve = 1

Probability distributions

- Binomial distribution
  - the number of successes in a sequence of n independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: success (with probability p) or failure (with probability $q = 1 - p$)
- $\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

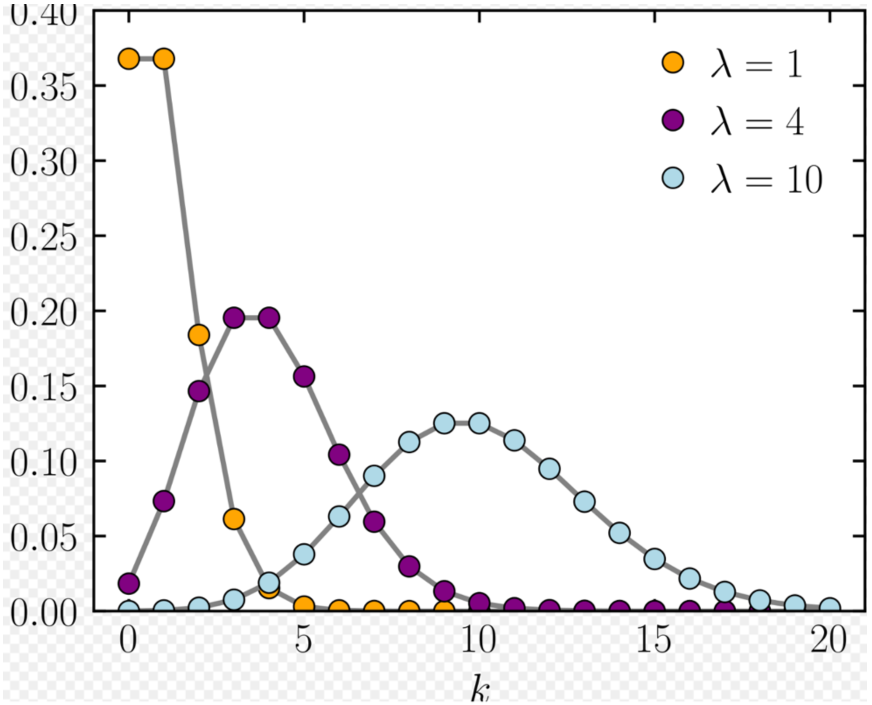
  $(k = 0, 1, \dots, n)$



- Mean: np
- Variance: npq

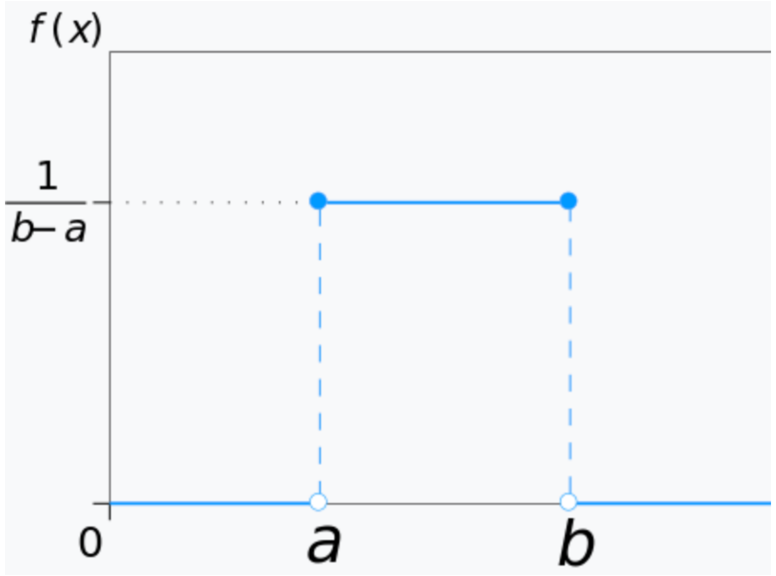
Probability distributions

- Poisson distribution
  - probability of a given number (k) of events occurring in a fixed interval of time or space
- $\Pr(X = k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$
- Mean: λ
- Variance: λ



Probability distributions

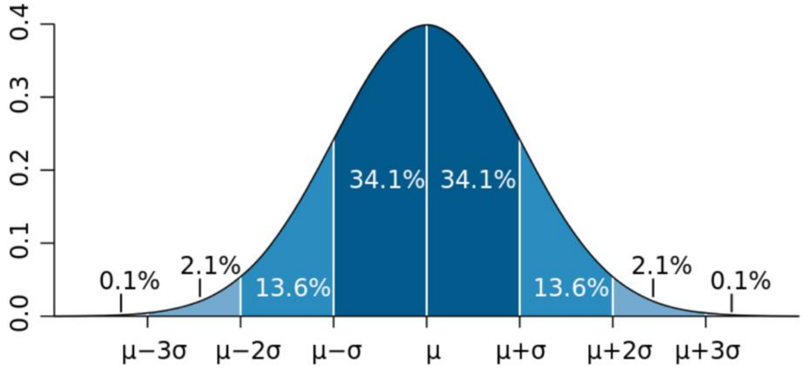
- Uniform distribution
- $f(x) = \frac{1}{b-a}, \;\; a \leq x \leq b$
- Mean: $(a + b)/2$
- Variance: $(b - a)^2/12$

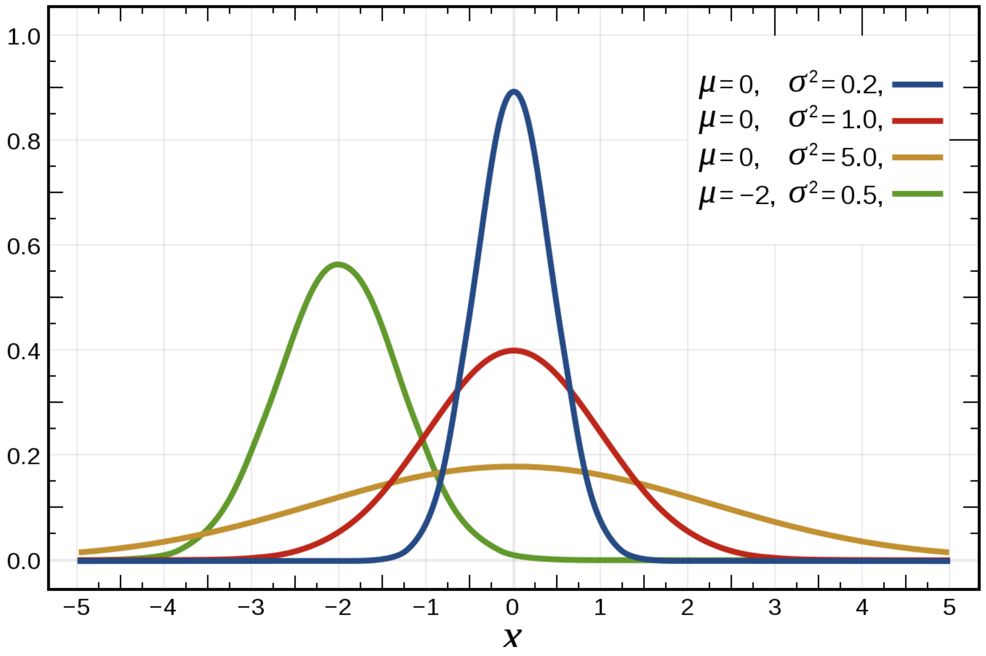


Probability distributions

- Normal distribution
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

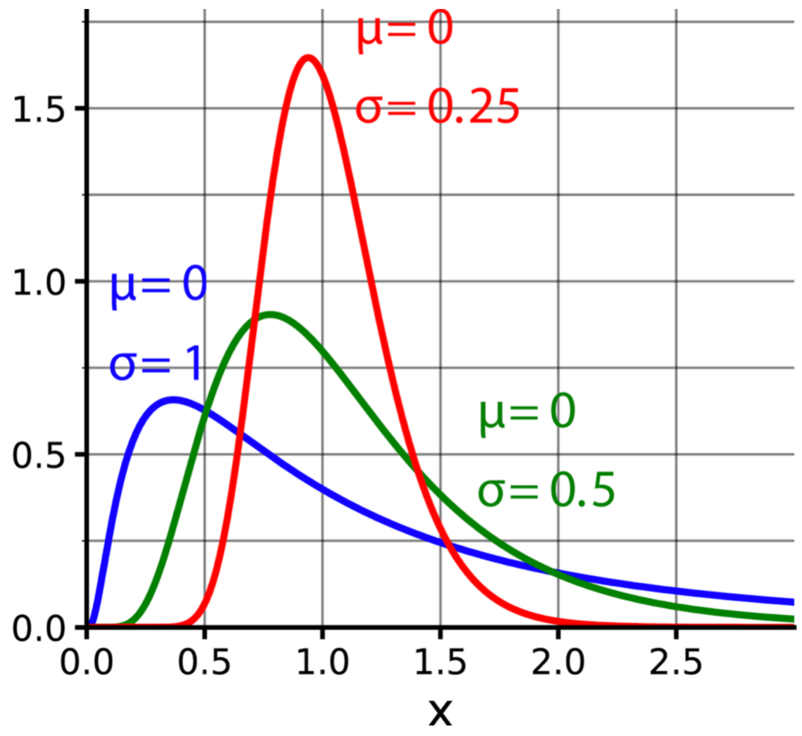
- Mean: μ
- Variance: σ^2





Probability distributions

- Lognormal distribution
  - if Y has a normal distribution, then $X = \mathrm{e}^{Y}$ has a lognormal distribution
- $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \; x > 0$

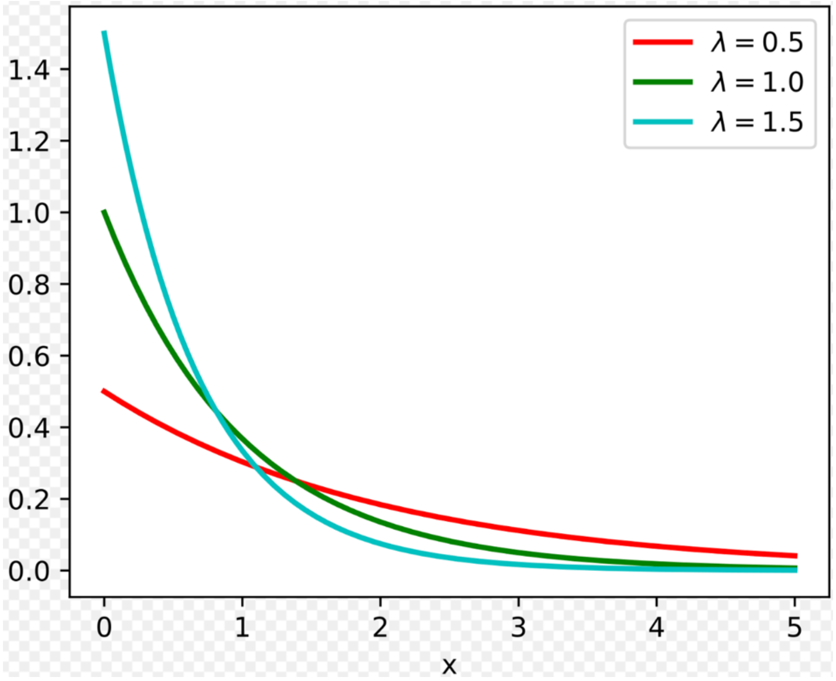


- Mean: $e^{\mu+\frac{\sigma^2}{2}}$
- Median: e^{μ}
- Variance: $(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$

Probability distributions

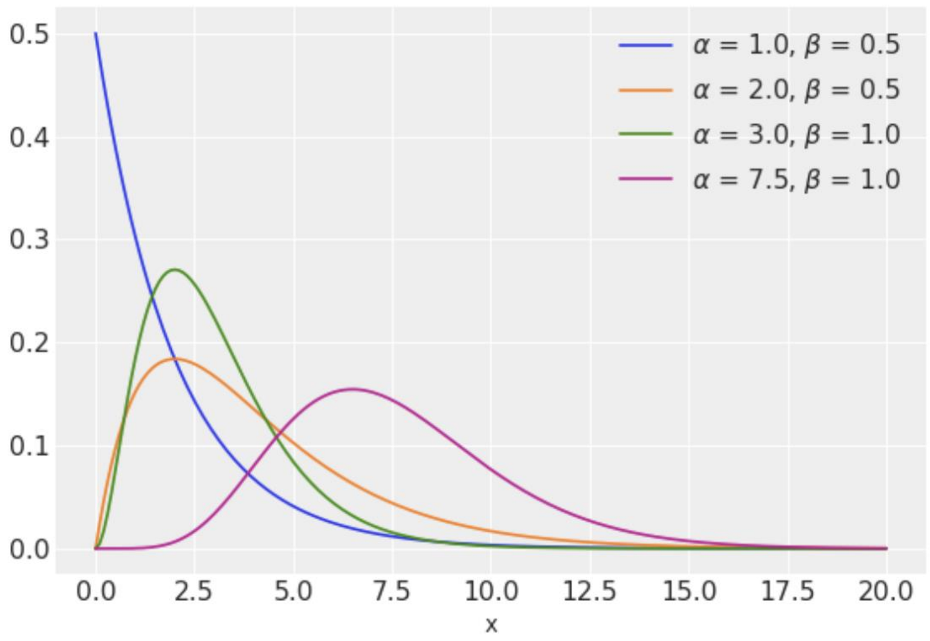
- Exponential distribution
- $f(x) = \lambda e^{-\lambda x},\ x \geq 0$

- Mean: $1/\lambda$
- Variance: $1/\lambda^2$



Probability distributions

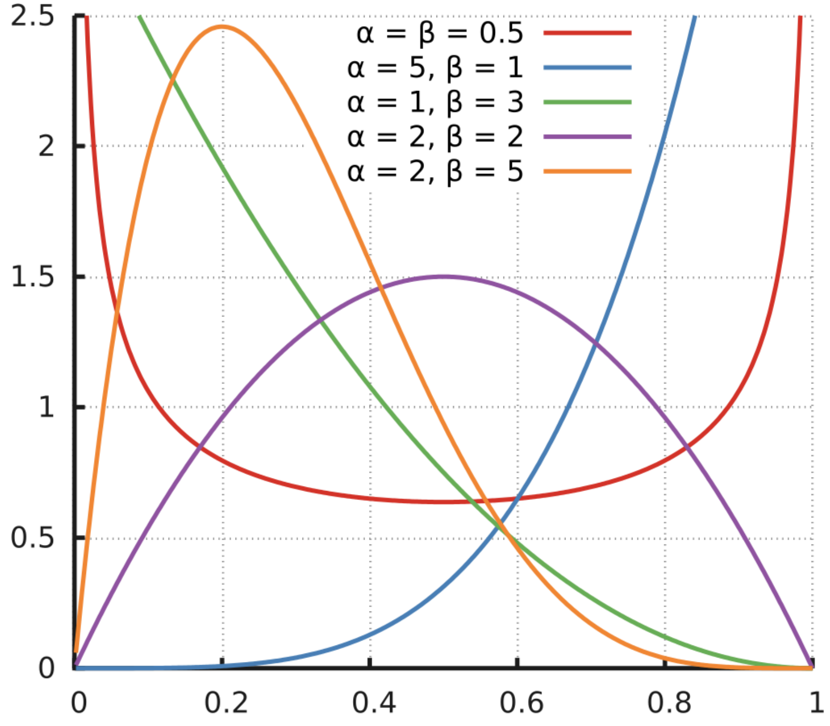
- Gamma distribution
- $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ x > 0$
- Mean: α/β
- Variance: α/β^2
- $\mathrm{Exp}(\beta)$ when $\alpha = 1$
- χ_n^2 when $\alpha = n/2$ and $\beta = 1/2$



Probability distributions

- Beta distribution
- $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},$

  $0 < x < 1$
- Mean: $\alpha/(\alpha + \beta)$
- Variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- $U(0, 1)$ when $\alpha = \beta = 1$



Probability distributions

- Dirichlet distribution
- $f(x_1, \dots, x_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$,

  $\sum_{i=1}^{K} x_i = 1, x_i \in [0,1]; i \in \{1, \dots, K\}$

- Mean: $\frac{\alpha_i}{\prod_{i=1}^{K} \alpha_i}$
- It is a multivariate generalization of the beta distribution