Gamma, Beta, and Dirichlet Distributions in Biological Sciences

Applications in Genetics, Phylogenetics, Ecology, and Bioinformatics

DRME

December 22, 2025

Introduction

Why These Distributions in Biology?

Key Characteristics that Make Them Suitable:

- **Gamma**: Models waiting times, mutation rates, branch lengths
- **Beta**: Models proportions, allele frequencies, probabilities
- **Dirichlet**: Models composition data, species abundances, gene expression
- All are **conjugate priors** in Bayesian analysis
- Support matches biological constraints
- Flexible shape parameters
- Mathematical tractability

Key Areas of Application

- **Genetics**: Allele frequencies, mutation rates, recombination
- **Phylogenetics**: Branch lengths, substitution models, tree priors
- **Ecology**: Species abundances, community composition, occupancy
- **Bioinformatics**: Sequence composition, gene expression, microbiome data

Genetics Applications

Gamma Distribution in Population Genetics

Application: Modeling Mutation Rate Variation

Problem: Mutation rates vary across genomic sites due to:

- Functional constraints
- Recombination rates
- Local GC content
- Chromatin structure

Solution: Use Gamma distribution to model rate variation:

$$r_i \sim \text{Gamma}(\alpha, \beta) \quad \text{for site } i$$

where r_i is the relative mutation rate at site i.

Interpretation:

- Shape parameter α controls rate heterogeneity
- Small α: High heterogeneity (few sites mutate fast)
- Large α: Low heterogeneity (rates more uniform)

Example: Gamma Distribution in dN/dS Analysis

dN/dS ratio: Measures selection pressure on protein-coding genes

Standard model: Single dN/dS ratio for all sites

Gamma model: Allow dN/dS variation across sites:

$$\omega_i \sim \text{Gamma}(\alpha, \beta) \quad \text{for codon site } i$$

Interpretation of parameters:

- $\omega_i < 1$: Purifying selection at site i
- $\omega_i = 1$: Neutral evolution at site i
- $\omega_i > 1$: Positive selection at site i

Result: Can identify:

- Proportion of sites under positive selection
- Specific sites with $\omega > 1$
- Variation in selective pressure

Application: Modeling Allele Frequencies

Wright-Fisher Model: Under genetic drift, allele frequency distribution follows:

$$f(p) = \frac{1}{B(2N_e\mu, 2N_e\nu)} p^{2N_e\mu - 1}(1-p)^{2N_e\nu - 1}$$

where:

- p: allele frequency
- N_e: effective population size
- μ: mutation rate to allele
- ν: mutation rate from allele

Beta Distribution in Allele Frequency Modeling

Application: Modeling Allele Frequencies

Interpretation: This is a Beta distribution:

$$p \sim \text{Beta}(2N_e\mu, 2N_e\nu)$$

Special cases:

- Neutral alleles: $\mu = \nu$, symmetric Beta
- Deleterious alleles: $\mu < \nu$, shifted toward 0

Example: Beta Distribution in GWAS

Genome-Wide Association Studies (GWAS): Identify SNPs associated with traits

Problem: Allele frequencies differ between cases and controls

Bayesian approach: Model allele frequencies with Beta priors:

$$\begin{aligned} \text{Controls:} \quad & p_{\text{control}} \sim \text{Beta}(\alpha_0, \beta_0) \\ \text{Cases:} \quad & p_{\text{case}} \sim \text{Beta}(\alpha_1, \beta_1) \end{aligned}$$

Example: Beta Distribution in GWAS

Interpretation:

- Prior parameters based on population data
- Update with observed genotype counts
- Compare posterior distributions
- Odds ratio: $\frac{p_{\text{case}}/(1-p_{\text{case}})}{p_{\text{control}}/(1-p_{\text{control}})}$

Advantage: Handles uncertainty, especially for rare variants

Dirichlet Distribution in Population Structure

Application: Admixture and Population Structure

ADMIXTURE model: Each individual's genome is mixture of ancestral populations

Model specification:

$$\mathbf{q}_i \sim \text{Dir}(\boldsymbol{\alpha}) \quad \text{for individual } i$$

where $\mathbf{q}_i = (q_{i1}, \ldots, q_{iK})$ are ancestry proportions from K populations.

Dirichlet Distribution in Population Structure

Interpretation:

- $\sum_{k=1}^{K} q_{ik} = 1$ (proportions sum to 1)
- α_k: Prior belief about contribution of population k
- Symmetric Dirichlet$(\alpha, \ldots, \alpha)$: No prior preference

Phylogenetics Applications

Gamma Distribution in Molecular Clock Models

Application: Rate Variation Among Lineages

Relaxed Molecular Clock: Allows evolutionary rates to vary among branches

Model: Each branch b has its own rate r_b:

$$r_b \sim \text{Gamma}(\alpha, \alpha)$$

constrained such that $E[r_b] = 1$

Gamma Distribution in Molecular Clock Models

Interpretation:

- α: Rate variation parameter
- Large α: Rates similar across branches (strict clock)
- Small α: High rate variation (relaxed clock)

Example: BEAST2 software implementation:

- `StrictClock`: $\alpha \to \infty$
- `RelaxedClockLogNormal`: Log-normal prior on rates
- `RelaxedClockExponential`: Exponential prior (Gamma(1,1))

Example: Gamma in Site Heterogeneity Models

GTR+Γ Model: General Time Reversible model with Gamma-distributed rates

For phylogenetics:

- Different sites evolve at different rates
- Fast-evolving sites: Less phylogenetic information
- Slow-evolving sites: More phylogenetic information

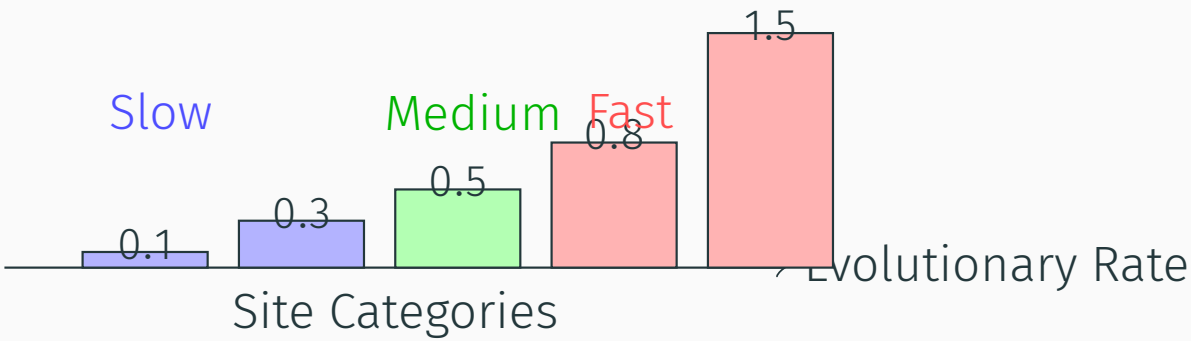
Model: Discrete Gamma approximation (Yang, 1994):

$$r_c \sim \text{Gamma}(\alpha, \alpha) \quad \text{discretized into } C \text{ categories}$$

Example: Gamma in Site Heterogeneity Models

GTR+Γ Model: General Time Reversible model with Gamma-distributed rates

Interpretation:



Beta Distribution in Tree Topology Priors

Application: Birth-Death Process for Tree Priors

Birth-Death Process: Models speciation and extinction

Parameterization:

- λ: Speciation rate
- μ: Extinction rate
- ρ: Sampling fraction

Tree prior probability:

$$P(T \mid \lambda, \mu, \rho) \propto \lambda^{n-2}(1 - \mu/\lambda)^m \prod_{i=1}^{n}(1 - p(t_i))^{x_i}$$

where $p(t)$ is solution to ODE and follows Beta distribution properties.

Beta prior on relative death rate:

$$d = \mu/\lambda \sim \text{Beta}(\alpha, \beta)$$

Beta Distribution in Tree Topology Priors

Application: Birth-Death Process for Tree Priors

Interpretation:

- $d \approx 0$: Little extinction (pure birth process)
- $d \approx 1$: High extinction relative to speciation

Dirichlet Distribution in Substitution Models

Application: Nucleotide Frequency Parameters

GTR Model: General Time Reversible substitution model

Parameters:

- $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$: Stationary frequencies
- $\sum \pi_i = 1$, $\pi_i > 0$

Natural prior: Dirichlet distribution

$$\boldsymbol{\pi} \sim \text{Dir}(\alpha_A, \alpha_C, \alpha_G, \alpha_T)$$

Dirichlet Distribution in Substitution Models

Application: Nucleotide Frequency Parameters

Common choices:

- Symmetric: $\text{Dir}(1, 1, 1, 1)$ (uniform)
- Empirical: $\text{Dir}(f_A + 1, f_C + 1, f_G + 1, f_T + 1)$ based on data
- Biological: $\text{Dir}(2, 1, 1, 2)$ if AT-rich expected

Interpretation: $\alpha_i - 1$ are "pseudo-counts" representing prior knowledge about base frequencies

Ecology Applications

Gamma Distribution in Species Abundance Models

Application: Modeling Species Abundances

Gamma distribution as prior for Poisson rate:

$$\lambda_i \sim \text{Gamma}(\alpha, \beta) \quad \text{for species } i$$

$$N_i \sim \text{Poisson}(\lambda_i) \quad \text{observed count}$$

This is Gamma-Poisson (Negative Binomial) model:

$$P(N_i = n) = \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)n!} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{n}$$

Gamma Distribution in Species Abundance Models

Application: Modeling Species Abundances

Interpretation:

- Accounts for overdispersion in ecological count data
- α: Shape (controls skewness of abundance distribution)
- β: Scale (controls mean abundance)
- Can model rare and common species simultaneously

Example: Gamma Distribution in Species-Area Relationships

Species-Area Relationship (SAR): $S = cA^z$

Problem: Parameter uncertainty and spatial correlation

Bayesian approach with Gamma priors:

$$c \sim \text{Gamma}(\alpha_c, \beta_c)$$
$$z \sim \text{Gamma}(\alpha_z, \beta_z)$$
$$S_i \sim \text{Poisson}(cA_i^z)$$

Example: Gamma Distribution in Species-Area Relationships

Interpretation of priors:

- c: Expected species richness at unit area
- z: Scaling exponent (typically 0.1-0.3)
- Gamma priors ensure positivity
- Can incorporate expert knowledge about likely values

Example values:

- Continental species: $z \sim \text{Gamma}(10, 50)$ (mean 0.2)
- Island species: $z \sim \text{Gamma}(15, 50)$ (mean 0.3)

Beta Distribution in Occupancy Modeling

Application: Species Presence-Absence Data

Occupancy model: Accounts for imperfect detection

Model structure:

$$
\begin{aligned}
z_i &\sim \text{Bernoulli}(\psi) \quad \text{true occupancy for site } i \\
y_{ij} \mid z_i = 1 &\sim \text{Bernoulli}(p) \quad \text{detection in survey } j \\
y_{ij} \mid z_i = 0 &= 0
\end{aligned}
$$

Beta priors for parameters:

$$
\begin{aligned}
\psi &\sim \text{Beta}(\alpha_\psi, \beta_\psi) \quad \text{occupancy probability} \\
p &\sim \text{Beta}(\alpha_p, \beta_p) \quad \text{detection probability}
\end{aligned}
$$

Beta Distribution in Occupancy Modeling

Application: Species Presence-Absence Data

Occupancy model: Accounts for imperfect detection

Interpretation:

- $\alpha - 1$: Prior "successes"
- $\beta - 1$: Prior "failures"
- Example: $\psi \sim \text{Beta}(2, 8)$ suggests occupancy 20
- Example: $p \sim \text{Beta}(8, 2)$ suggests detection 80

Dirichlet Distribution in Community Composition

Application: Microbial Community Data

Microbiome data: Counts of different taxa in samples

Dirichlet-Multinomial model:

$$\mathbf{p}_i \sim \text{Dir}(\boldsymbol{\alpha}) \quad \text{for sample } i$$

$$\mathbf{y}_i \sim \text{Multinomial}(N_i, \mathbf{p}_i) \quad \text{observed counts}$$

Dirichlet Distribution in Community Composition

Application: Microbial Community Data

Interpretation:

- $\mathbf{p}_i$: True proportions of taxa in sample i
- $\boldsymbol{\alpha}$: Controls overall composition and variability
- $\alpha_0 = \sum \alpha_j$: Concentration parameter
- Small α_0: High between-sample variability
- Large α_0: Similar composition across samples

Advantages over Multinomial:

- Accounts for overdispersion
- Models correlation between taxa
- Allows borrowing information across samples

Example: Dirichlet in Beta Diversity Analysis

Beta diversity: Differences in community composition between samples

Dirichlet prior for sample-specific compositions:

$$\mathbf{p}_i \sim \text{Dir}(\boldsymbol{\alpha}_i)$$

where $\boldsymbol{\alpha}_i$ depends on environmental covariates:

$$\log(\alpha_{ij}) = \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} X_{ik}$$

Interpretation:

- β_{kj}: Effect of covariate k on taxon j
- Positive β: Higher proportion with increasing covariate
- Negative β: Lower proportion with increasing covariate

Example: Dirichlet in Beta Diversity Analysis

Beta diversity: Differences in community composition between samples

Interpretation:

- β_{kj}: Effect of covariate k on taxon j
- Positive β: Higher proportion with increasing covariate
- Negative β: Lower proportion with increasing covariate

Bioinformatics Applications

Example: Gamma in Sequence Alignment Scores

Application: BLAST E-value Calculation

Extreme value distribution for alignment scores: Gumbel distribution

For local alignment without gaps (Karlin-Altschul theory):

$$P(S > x) \approx 1 - \exp(-Kmne^{-\lambda x})$$

where parameters K, λ estimated from data.

Gamma approximation for sum of scores: For multiple hits

$$\text{Sum of scores} \sim \text{Gamma}(n, \lambda)$$

where n is effective number of independent hits.

Example: Gamma in Sequence Alignment Scores

Application: BLAST E-value Calculation

Interpretation in BLAST:

- E-value: Expected number of hits with score $\geq S$
- $E = mn2^{-S}$ (simplified form)
- Small E-value: Significant alignment
- Large E-value: Likely by chance

Comparative Analysis

Comparing Applications Across Fields

Distribution	Genetics	Phylogenetics	Ecology	Bioinformatics
Gamma	Mutation rate variation dN/dS variation	Site rate heterogeneity Branch length variation	Species abundances Species-area curves	RNA-seq dispersion Alignment scores
Beta	Allele frequencies GWAS priors	Tree prior parameters Relative extinction rates	Occupancy probabilities Detection probabilities	Genotype likelihoods Methylation levels
Dirichlet	Ancestry proportions Haplotype frequencies	Base frequencies Substitution model	Community composition Species proportions	Taxonomic assignment Pathway abundances

Comparing Applications Across Fields

Common Themes:

- All handle positive-valued parameters
- Conjugate priors enable Bayesian computation
- Flexible shapes accommodate diverse biological patterns
- Hierarchical modeling across biological units

Implementation in some Software Packages

Popular Tools Using These Distributions:

Genetics/Phylogenetics:

- **BEAST2**: Gamma for clock models, Dirichlet for base frequencies
- **(Mr|Rev)Bayes**: Gamma for rate variation, Beta for tree parameters
- **ADMIXTURE**: Dirichlet for ancestry proportions
- **GATK**: Beta-binomial for variant calling

Ecology/Bioinformatics:

- **DESeq2**: Gamma-Poisson for RNA-seq
- **Phyloseq**: Dirichlet-multinomial for microbiome
- **MEGAN**: Dirichlet for taxonomic assignment

R/Python Packages:

- R: `MCMCpack`, `dirichlet`, `rstan`, `brms`
- Python: `pymc3`, `tensorflow-probability`, `scipy.stats`
- Julia: `Turing.jl`, `Distributions.jl`
- R/Python: `tppl[?]`

Choosing Appropriate Distributions:

When to Use Gamma

- Modeling rates, waiting times, or positive continuous variables
- When you need flexible right-skewed distributions
- As prior for precision in Normal models
- When data show overdispersion relative to Poisson

Choosing Appropriate Distributions:

When to Use Beta

- Modeling proportions, probabilities, or percentages
- When variable is bounded between 0 and 1
- As prior for success probability in Binomial models
- When you want interpretable parameters (pseudo-counts)

Choosing Appropriate Distributions:

When to Use Dirichlet

- Modeling vectors of proportions that sum to 1
- Compositional data (species, genes, etc.)
- As prior for Multinomial probabilities
- When you need to model correlations between proportions

Challenges and Solutions

Common Challenges in Biological Applications:

Computational Challenges:

- High-dimensional Dirichlet distributions
- Slow MCMC convergence
- Large datasets (millions of observations)
- Model identifiability issues

Challenges and Solutions

Solutions:

- Use variational inference approximations
- Implement in Stan/Nimble for efficient sampling
- Use sparse representations
- Add regularization/priors

Challenges and Solutions

Biological Challenges:

- Zero-inflation (many taxa absent)
- Phylogenetic non-independence
- Batch effects and technical noise
- Missing data and censoring

Challenges and Solutions

Solutions:

- Zero-inflated models
- Phylogenetic correlation structures
- Include batch covariates
- Hierarchical modeling

Emerging Approaches:

- Deep learning with probabilistic layers
- Gaussian process extensions
- Scalable variational inference
- Integrated models combining multiple data types

Future Directions

Emerging Applications and Methods:

New Biological Questions:

- Single-cell multi-omics integration
- Spatial transcriptomics and proteomics
- Microbiome-host interaction networks
- Time-series and longitudinal data
- Cross-species comparative analysis

Future Directions

Methodological Advances:

- Neural density estimation
- Differentiable probabilistic programming
- Federated learning for privacy
- Causal inference with compositional data
- Transfer learning across studies

Future Directions

Key Resources for Practitioners:

- **Books**: Bayesian Data Analysis (Gelman et al.), Statistical Rethinking (McElreath)
- **Courses**: Statistical Genetics, Ecological Statistics, Computational Biology
- **Journals**: MBE, Bioinformatics, Ecology, Molecular Ecology