Generalized linear models:

General linear models +

Non-normal residuals=

Generalized linear models

Negative binomial regression,
beta regression and more…

Generalized linear models

Main ideas:

1) A *transformation of the expectation* of the response is expressed as a linear combination of covariates rather than the mean response directly.
2) For the random part of the model, *distributions other than the Normal* can be chosen, e.g., Poisson, Binomial or gamma.

Generalized linear models

GLMs are made up of three components:

1) A *statistical distribution* used to describe the random variation in the response *y*; this is the stochastic part of the system description
2) A *linear predictor*, i.e., a linear combination of covariate effects that are thought to make up *E(y)*; this is the systematic or deterministic part of the system description.
3) A *link function* that is applied to the *E(y)*, expectation of the response

Objectives

- So far, we've modeled data with 3 distributions: normal, Poisson, binomial
- Discuss modeling data with other distributions, namely the negative binomial and beta distributions
  - Understand when to use each distribution

GLMs …. So far

Distribution	Dependent data type	When to use it?
Normal	Continuous $(-\infty, \infty)$	In cases where errors around means are normal
Poisson	Discrete positive values $(0, \infty)$	Modeling counts under the assumption that mean=variance
Binomial	Discrete positive values with an upper bound (0, N)	Modeling the probability of C successes in N trials

Introducing the Gamma distribution

Gamma distribution

A *continuous* distribution with a positive support – exponential and $\boldsymbol{\chi}^2$ are special cases

- Denoted: $x \sim \Gamma(k, \theta) \quad k, \theta > 0$
- Mean: $k\theta$
- Variance: $k\theta^{\,2}$
- Support: $x \in (0, \infty)$

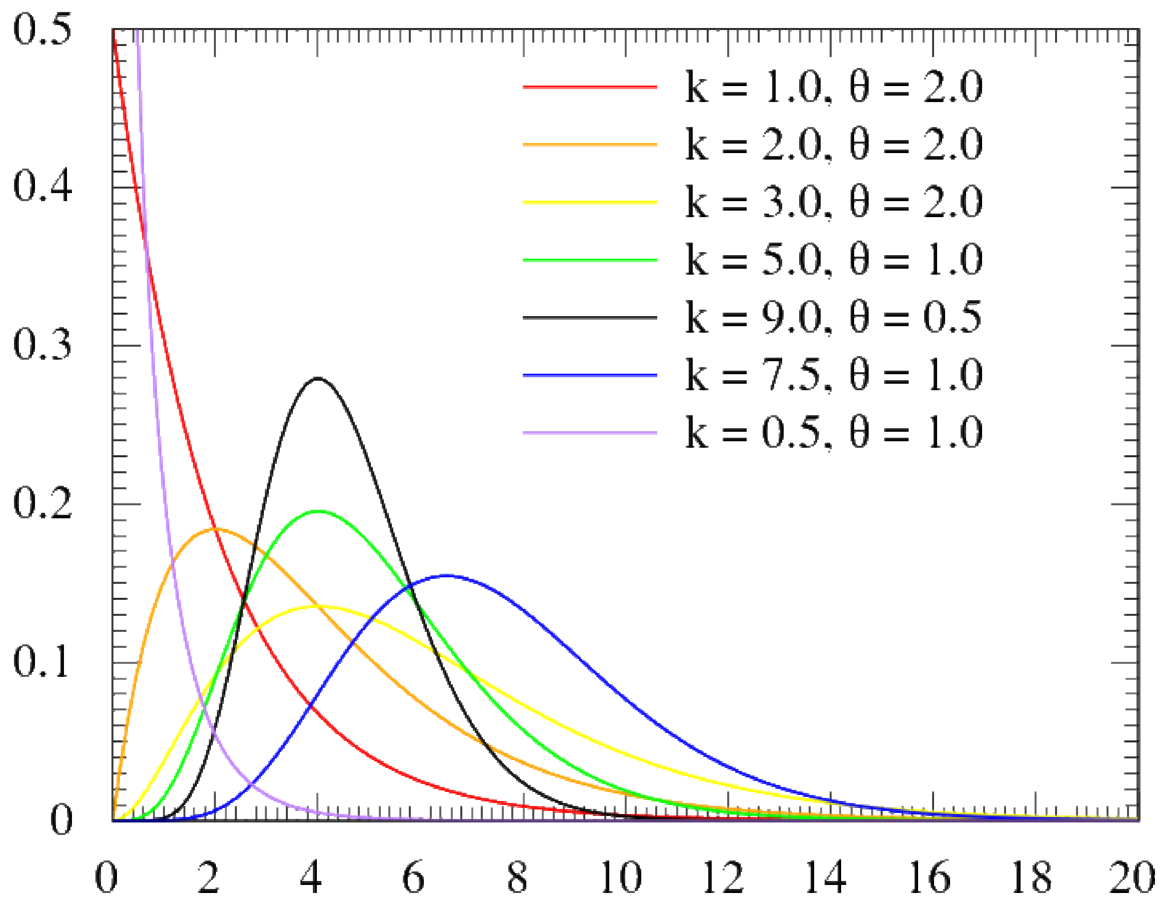
Gamma distribution

Probability density function:

$$p(x \mid k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

Where

$$\Gamma(k) = (k-1)! = \int_0^\infty x^{k-1} e^{-x} dx$$



There might be an instance when you want to model something according to the gamma…

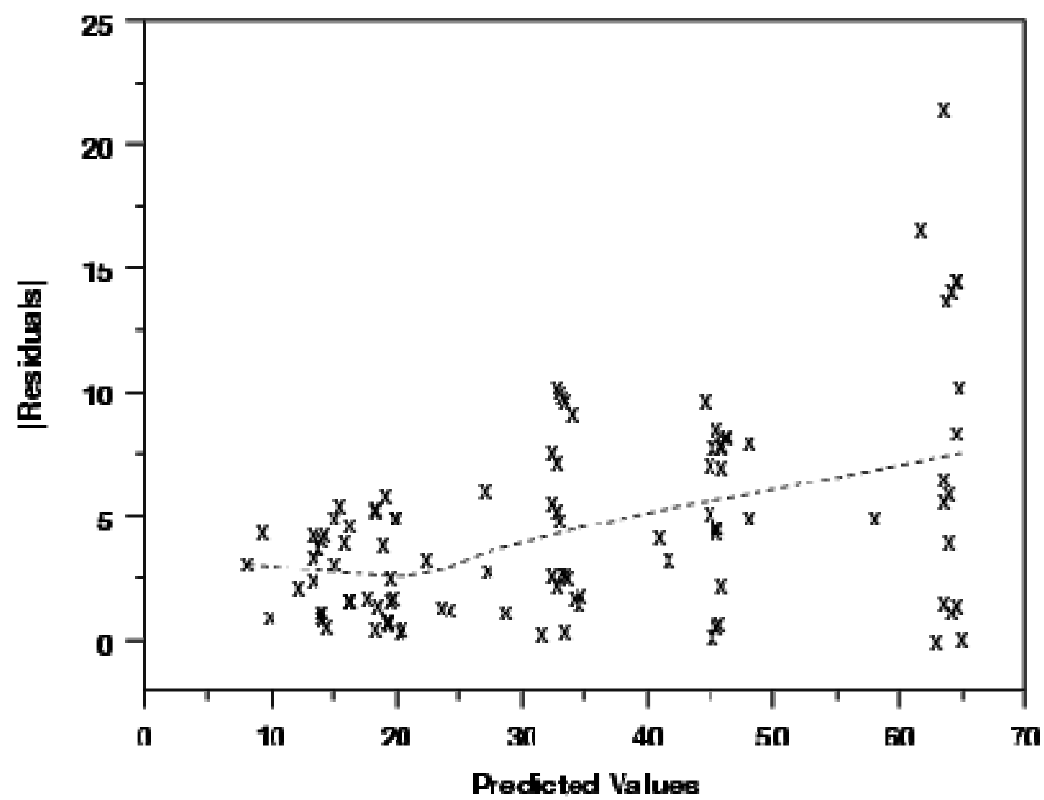
- In the instance of a continuous positive value that is heavily skewed, such that the error around the residuals is not normal

There might be an instance when you want to model something according to the gamma… but it doesn't occur that often

- However, the gamma distribution is useful for understanding other kinds of modeling
- Also useful as a prior in Bayesian analyses

Negative binomial regression

- Poisson distribution commonly used for modeling counts but…. mean=variance assumption can be quite restrictive
- In practice, most data won't mean this assumption
- Data are frequently ***overdispersed***



GLMs: stochastic part

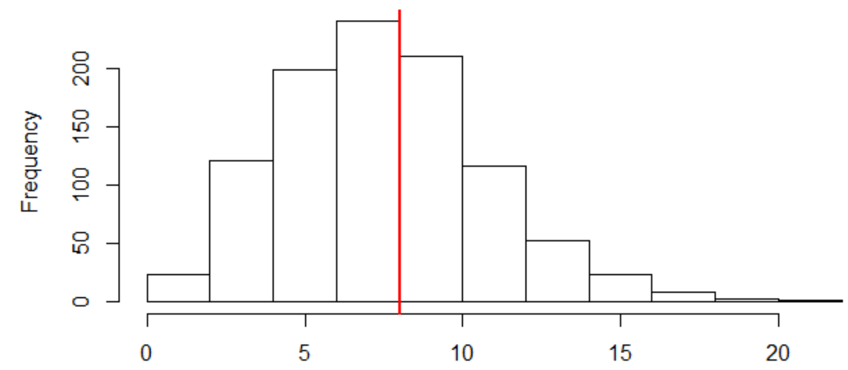
Negative binomial Distribution

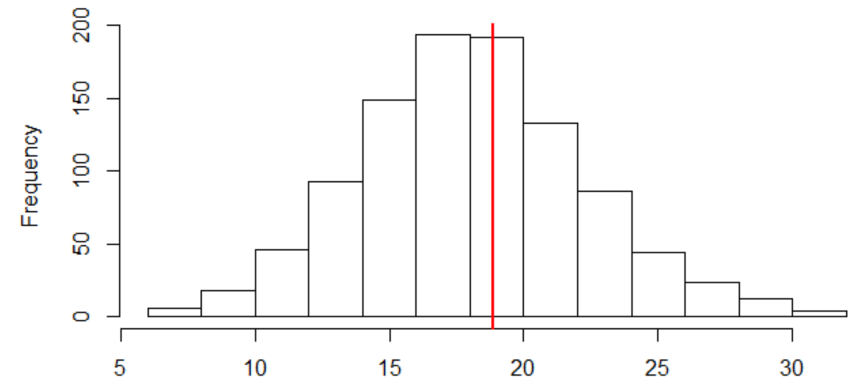
An important *discrete* distribution that is useful for modeling counts

- Denoted: $x \sim NB(r, p)$
- Mean: $r(1-p)/p$
- Variance: $r(1-p)/p^2$
- Support: $x \in \{0,1,2,3, \dots\}$

GLMs: stochastic part
Negative binomial Distribution

- *Classical examples:* Number of individuals in a flock of birds or school of fish. (e.g., small probabilities of large values)
- *Varieties:* Mixture of the Poisson and gamma distributions (i.e., an overdispersed Poisson)
- *Mathematical description*: 2 parameters: success probability p, and "size", r. Consider a number of Bernoulli trials in which the probability of success = p. Observe this sequence until a predefined number *r* of failures. The number of successes, x, has a negative binomial distribution.



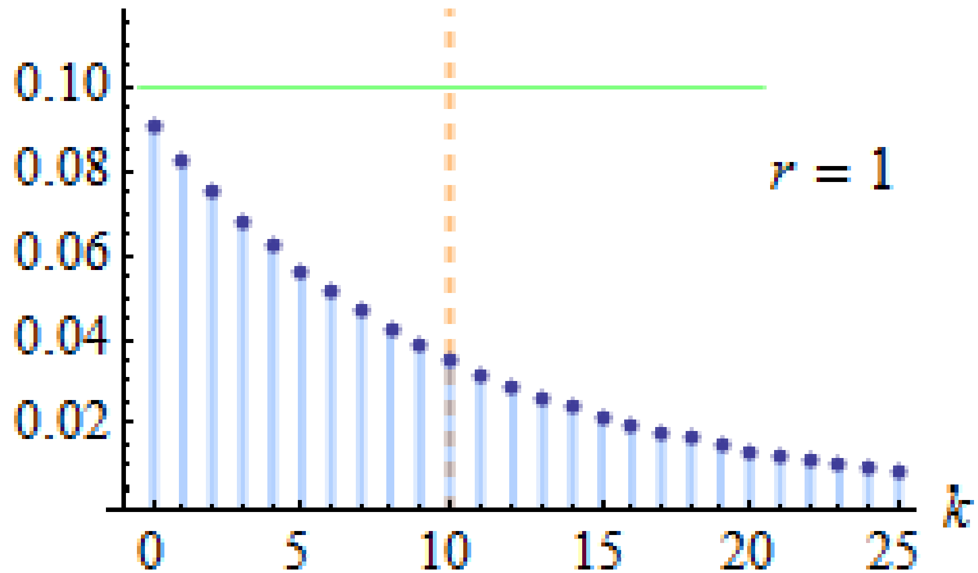


GLMs: stochastic part

Negative binomial Distribution

Probability mass function:

$$p(x = k \mid r, p) = \frac{(k + r - 1)!}{k!\,(r - 1)!} p^r (1 - p)^k$$



Likelihood function:

$$L(r, p \mid X) = \prod_{i}^{n} \frac{(x_i + r - 1)!}{x_i!\,(r - 1)!} p^r (1 - p)^{x_i}$$

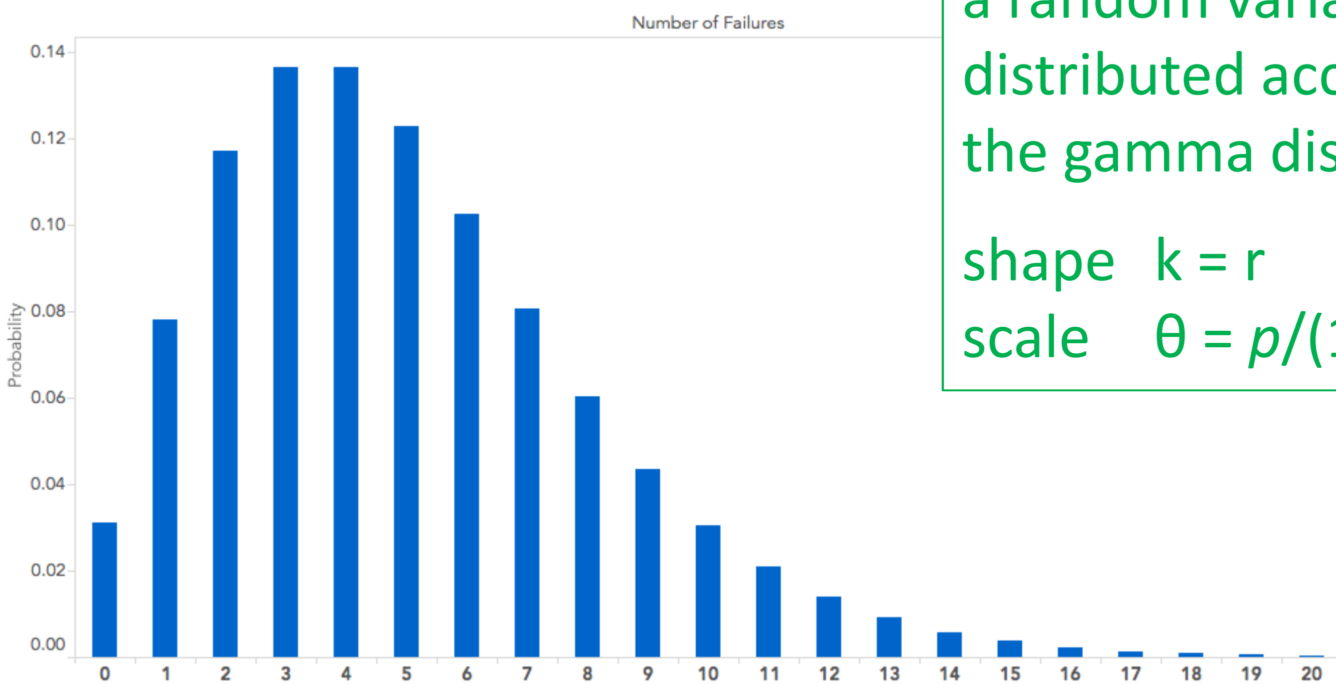
Where $X = \{x_1, x_2, \ldots, x_n\}$

GLMs: stochastic part

Negative binomial Distribution

Probability mass function:

$$p(x = k \mid r, p) = \frac{\Gamma(k + r)}{\Gamma(r)k!} p^r (1 - p)^k$$



Can view the negative binomial as a Pois(λ) distribution where λ is a random variable distributed according to the gamma distribution:

shape $k = r$
scale $\theta = p/(1 - p)$

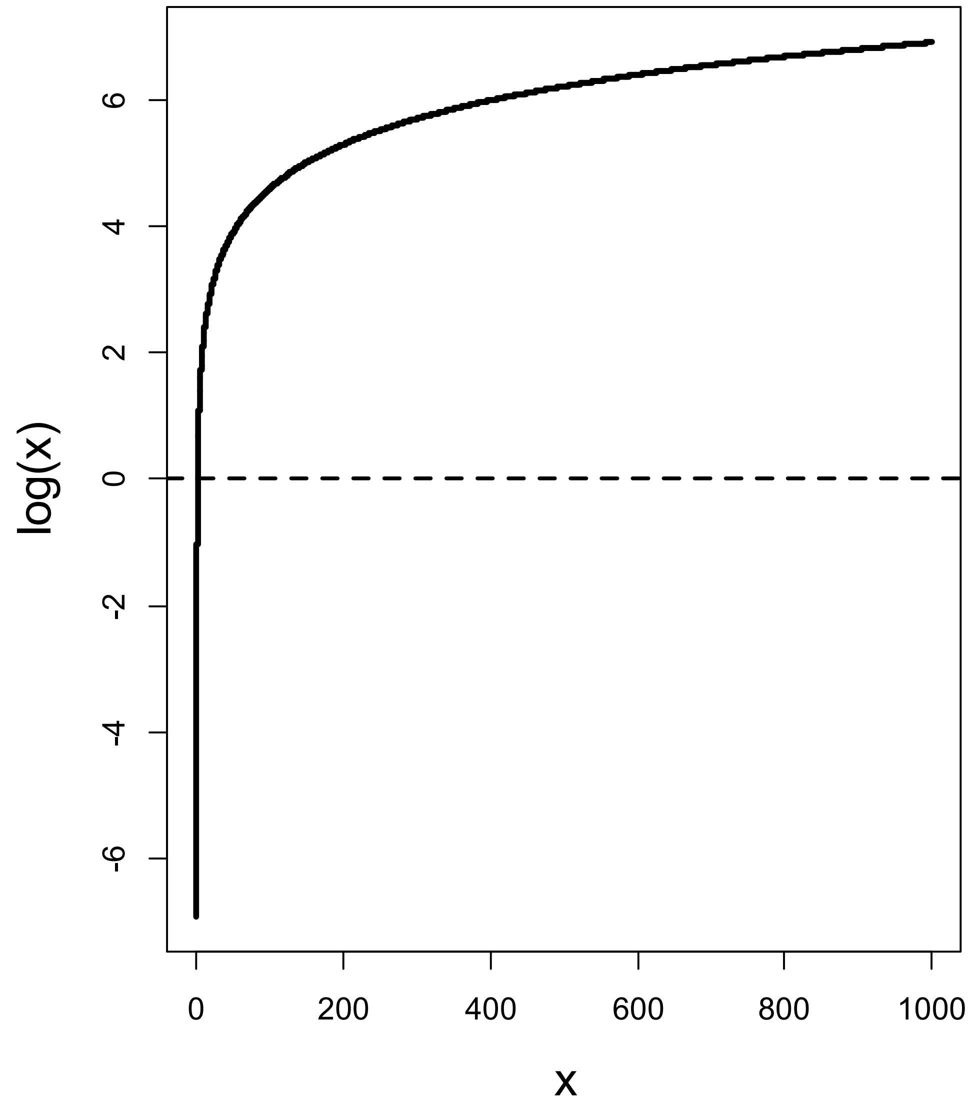
Which link function to use?

- Modeling counts (whole numbers)
- Expected count must be greater than or equal to zero

Log(x)

- Range: $x > 0$
- Range:

  $-\infty > \log(x) > \infty$



Negative binomial regression

Estimate the abundance of sea lions relative to local human abundances and amount of prey.

- Data collection:
  - n_i the number of sea lions counted at location *i*.
  - $human_i$ = people per sq km within 100 km of location *i*.
  - $prey_i$ = the estimated amount of available prey at *i*.

Is human presence negatively correlated with sea lion abundance?

Negative binomial regression

What is the distribution, link function, and linear predictor we should use?

Distribution:

Link function:

Linear predictor:

Negative binomial regression

What is the distribution, link function, and linear predictor we should use?

Distribution: $n_i \sim NB(r_i, p_i)$

Link function:

Linear predictor:

Negative binomial regression

What is the distribution, link function, and linear predictor we should use?

Distribution: $n_i \sim NB(r_i, p_i)$

Link function: $\log(\mu_i) = \log(r_i(1 - p_i)/p_i)$

Linear predictor:

Negative binomial regression

What is the distribution, link function, and linear predictor we should use?

Distribution: $n_i \sim NB(r_i, p_i)$

Link function: $\log(\mu_i) = \log(r_i(1 - p_i)/p_i)$

Linear predictor:

Note: We want to model the covariates relative to μ, the mean, (on the log scale) not r or p.

Negative binomial regression

What is the distribution, link function, and linear predictor we should use?

Distribution: $n_i \sim NB(r_i, p_i)$

Link function: $\log(\mu_i) = \log(r_i(1 - p_i)/p_i)$

Linear predictor: $\alpha + \beta 1 * human_i + \beta 2 * prey_i$

Note: We want to model the covariates relative to μ, the mean, (on the log scale) not r or p.

Moving on to the

beta distribution…..

Beta regression

- Suppose you want to model proportions (e.g., % forest cover relative to tree basal area; allele frequencies)
- How should we one perform a regression analysis in which the dependent variable is restricted to the standard unit interval such as ***rates*** and ***proportions***?

GLMs: stochastic part

Beta Distribution

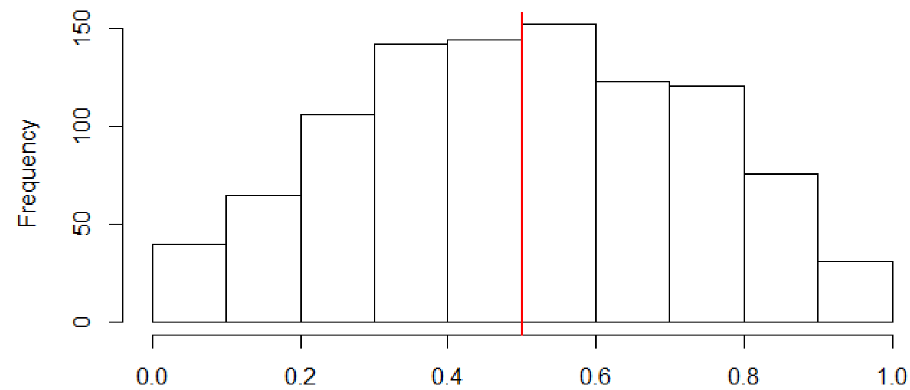
A *continuous* distribution that is useful for modeling values between 0 and 1

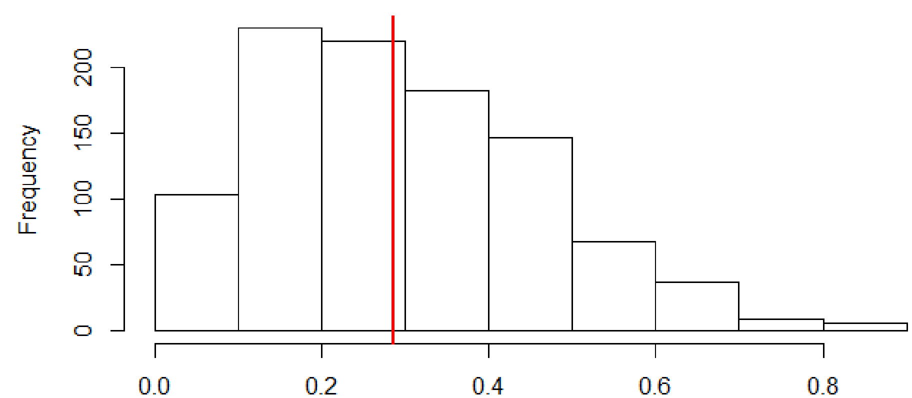
- Denoted: $Beta(a, b) \quad a, b > 0$
- Mean: $a/(a + b)$
- Variance: $ab/[(a + b)^2(a + b + 1)]$
- Support: $x \in (0,1)$

GLMs: stochastic part
Beta Distribution

- *Sampling situation:* Modeling the random behavior of percentages, rates, and proportions
- *Classical examples:* Allele frequency; Genetic distance between two populations; variability of soil properties; site connectivity
- *Varieties:* 1) The continuous uniform distribution between 0 and 1 is a special case where a = b = 1 ;

  2) The Balding–Nichols model is an alternative parametrization used in population genetics.
- *Mathematical description*: 2 shape parameters: a and b, both greater than zero





GLMs: stochastic part

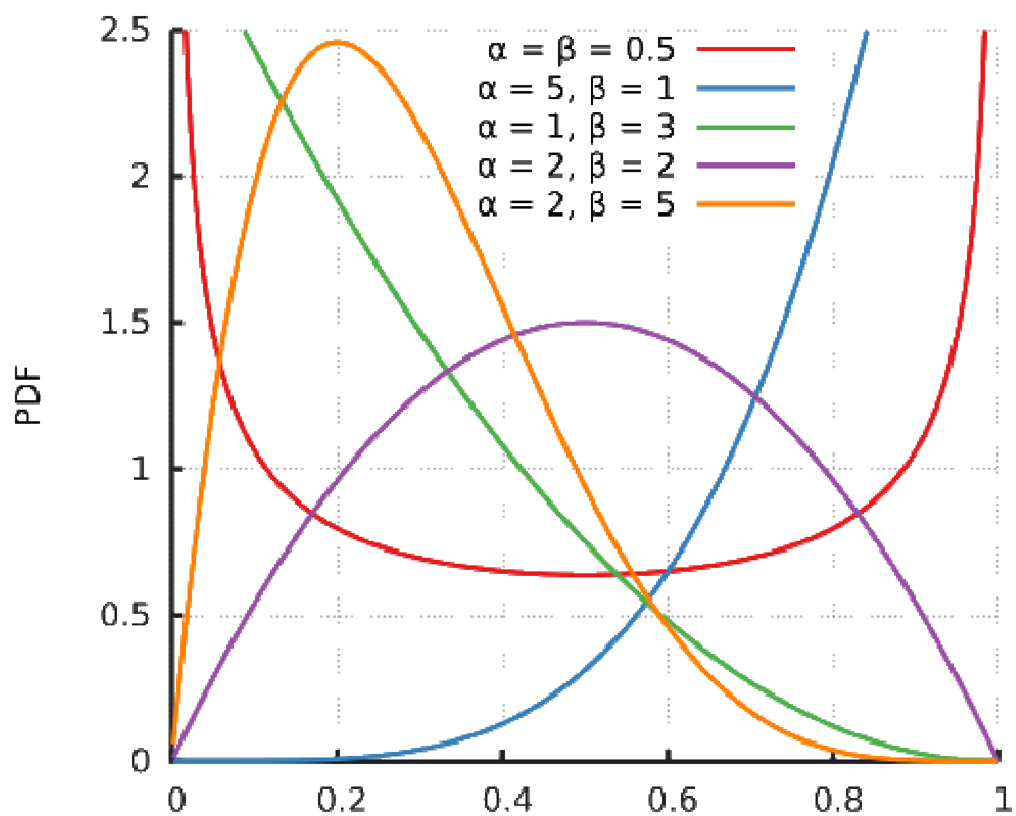
Beta Distribution

Probability density function:

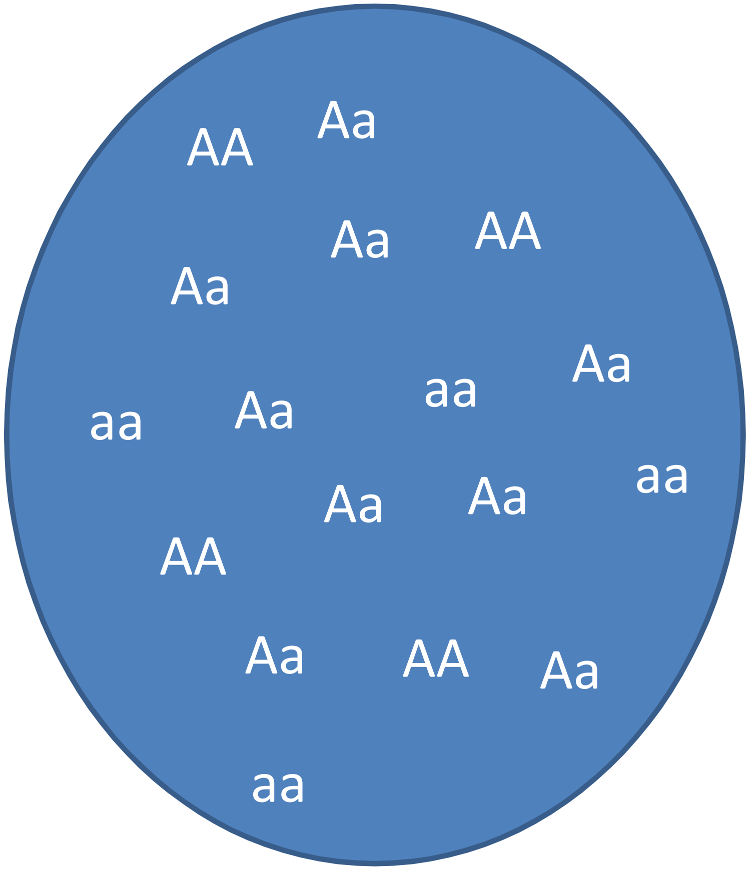
$$p(x \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

Where

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$$



Beta t-test



Estimate whether there is a difference in gene frequency (proportion of aa) in populations that come from one of two ancestors.

- Data collection:
  - y_i the proportion of aa genotypes in population *i*.
  - $ancestor_i$ = is an indictor vector where the value = 0 if the ancestor was from the first ancestor or 1 if from the second for population *i*.

Beta t-test

What is the distribution, link function, and linear predictor we should use?

Distribution:

Link function:

Linear predictor:

Beta t-test

What is the distribution, link function, and linear predictor we should use?

Distribution: $y_i \sim Beta(a_i, b_i)$

Link function:

Linear predictor:

Beta t-test

What is the distribution, link function, and linear predictor we should use?

Distribution: $y_i \sim Beta(a_i, b_i)$

Link function: $\mathrm{logit}\,(\mu_i)$

Linear predictor:

Beta t-test

What is the distribution, link function, and linear predictor we should use?

Distribution: $y_i \sim Beta(a_i, b_i)$

Link function: $\text{logit}\ (\mu_i)$

Linear predictor: $\alpha + \beta * ancestor_i$

Beta t-test

What is the distribution, link function, and linear predictor we should use?

Distribution: $y_i \sim Beta(a_i, b_i)$

Link function: $\text{logit}\ (\mu_i)$

Linear predictor: $\alpha + \beta * ancestor_i$

Note: We want to model the covariates relative to μ, the mean, (on the logit scale) not a or b.

$$logit\ (\mu_i) = logit\ (a_i/(a_i + b_i))$$

Generalized linear models

Lab: Negative binomial regression