Probability Distributions Part II

IBIO 851

Sept 22 2016

Suggested reading: **Chapter 4** in **Ecological Models & Data in R** (Bolker)

Some clarifications

- Pbinom() command clarification
- In class code, I said that pbinom takes 3 arguments: # of successes, # of trials, probability
- Actually: it's AT LEAST # of successes, # trials, probability
  - There's a default lower.tail=TRUE argument I didn't mention in the code
- pbinom(25, 50, 0.5) → what's the probability of 25 or fewer successes given 50 trials and a per-trial probability of 0.5

Some clarifications

- Pbinom() & pnorm() clarification
- Lower.tail=TRUE gives probabilities ≤ x (<u>including</u> # you specify)
- Lower.tail=FALSE gives probabilities > x (<u>not</u> including number you specify)
- If you want: $P(45 < x < 55)$ for x~bin(100,0.5), do:
  - pbinom(54, 100, 0.5, lower.tail=TRUE) - pbinom(46, 100, 0.5, lower.tail=TRUE)

Some clarifications

- What parameter are we estimating with a binomial distribution?
- In the case of the fly example from last class, you were getting at the # of successful trials
- But in most cases (i.e., statistical modeling), you're estimating the proportion of successes in the entire population

Why care about probability distributions?

- You can use a given probability distribution to describe a particular variable
- The distribution encapsulates our knowledge about the value of the random variable we're interested in
- Each probability distribution is characterized by certain parameters, some of which you're interested in estimating with a model

Goals for today

- Continue with discrete probability distributions
- Go over continuous probability distributions

Poisson distribution

- Evaluates the probability of a (usually small) number of occurrences out of many opportunities in a …
  - Period of time
  - Area
  - Volume
  - Weight
  - Distance
  - Other units of measurement
- Useful for counting rare events like new migrants to a population/year

Poisson distribution

Probability mass function:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- λ = mean number of occurrences in the given unit of time, area, volume, etc. (i.e. rate parameter)
- e = 2.71828….
- In Poisson distributions, the mean = variance
  - $\mu = \lambda$
  - $\sigma^2 = \lambda$
- Only 1 parameter!

Poisson distribution

- Let's say in a given stream there are an average of 3 striped trout per 100 yards. What is the probability of seeing 5 striped trout in the next 100 yards, assuming a Poisson distribution?

$$P(x=5)=\frac{\lambda^{x}e^{-\lambda}}{x!}=$$

Poisson distribution

- How about the next 50 yards?
  - Since the distance is only half as long, λ is only half as large

$$P(x=5)=\frac{\lambda^x e^{-\lambda}}{x!}=$$

Poisson distribution: example

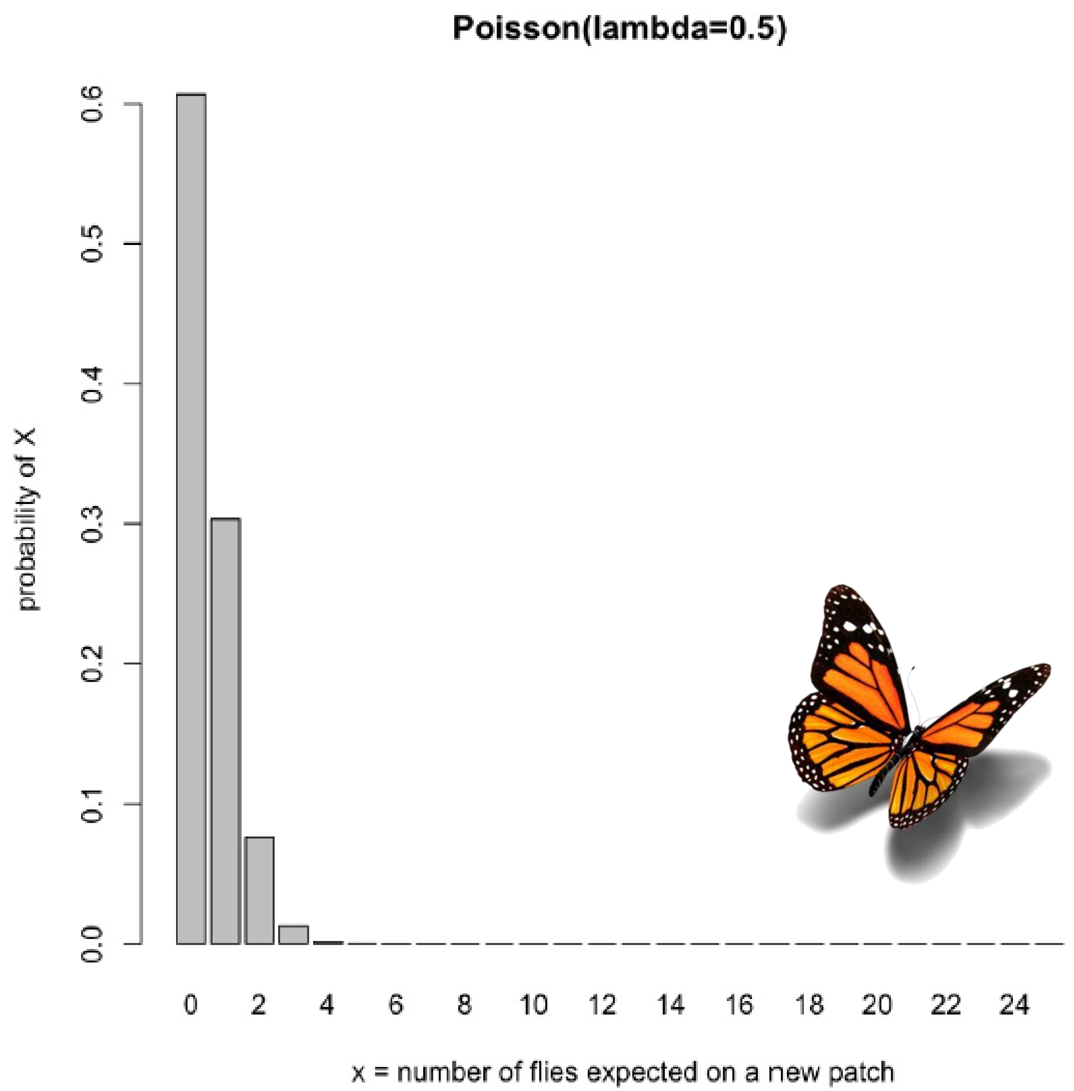
- Say monarch butterflies disperse to colonize a new patch at a very low rate (previous estimates suggest we will observe one butterfly for every 2 patches we examine, $\lambda = 0.5$)
- What is the probability of observing 2 butterflies on a new patch of land?

Poisson distribution: example

$$P(x=2) = \frac{\lambda^x e^{-\lambda}}{x!} =$$

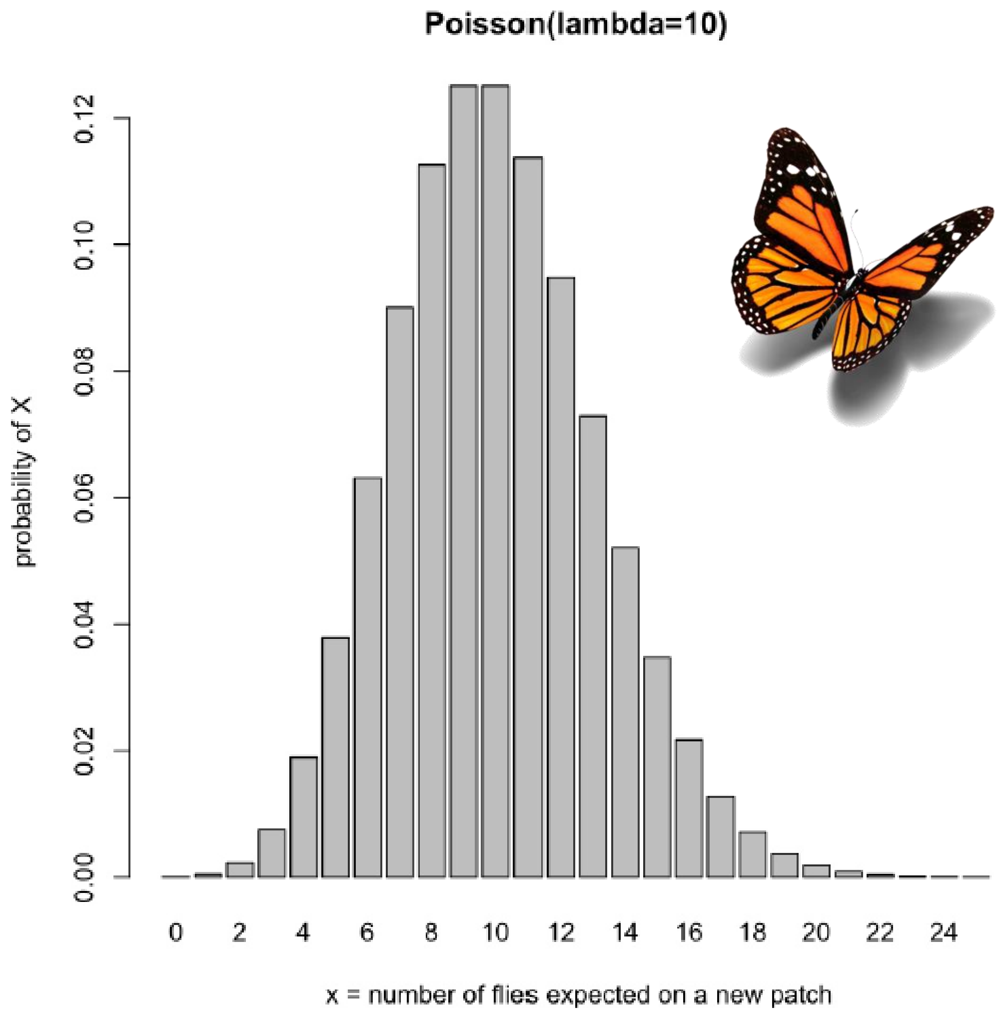
We have ~8% chance of seeing 2 butterflies at the next patch of land

Poisson distribution: example



Poisson distribution: example

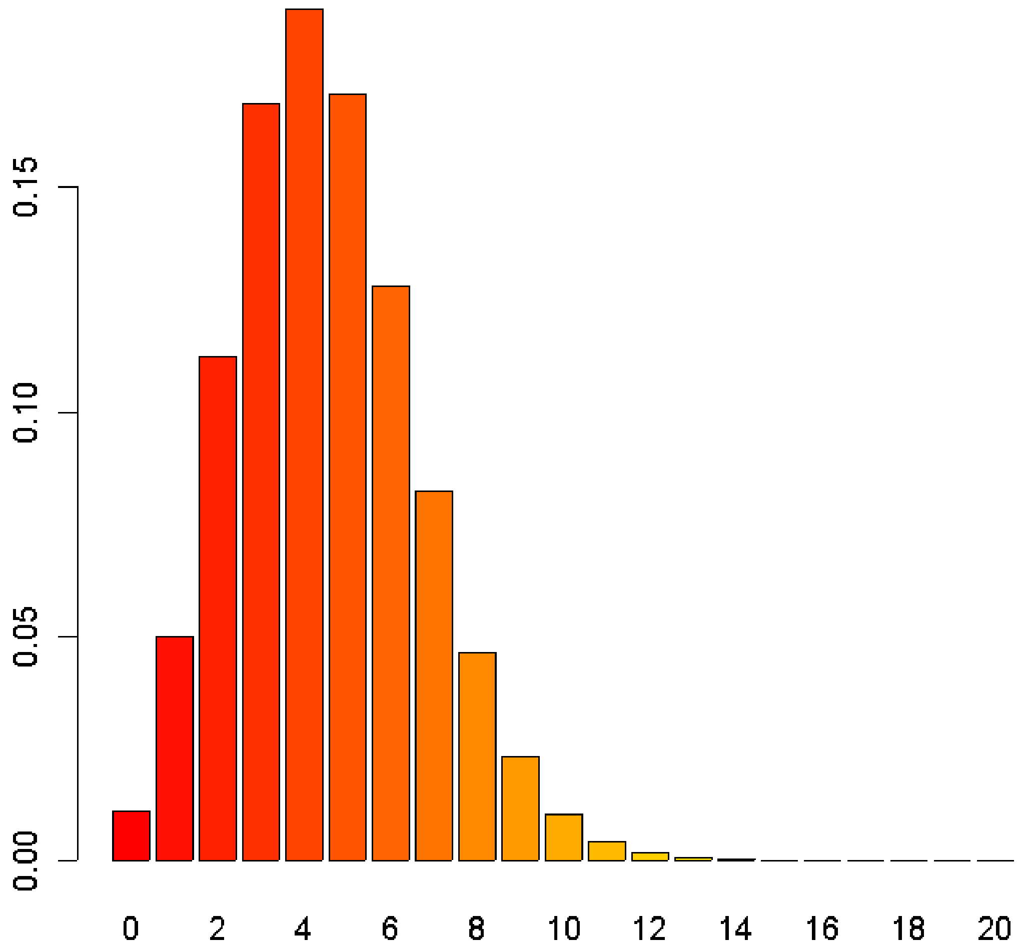
- What happens as λ increases?



Poisson distribution in R

- As with the binomial distribution, the codes:
  - dpois & ppois will do the calculations for density & probability for you
- Specify your vector of quantiles & lambda
- X <- dpois(0:20, 4.5) &
- Barplot(X) would get you:

Poisson distribution in R



Other Poisson examples

- Number of offspring in a season
- Number of prey caught per unit of time
- Number of migrants passing overhead at a watch station
- **Key point:** Use when binomial doesn't work because of the rare occasion of large counts

When Poisson doesn't quite fit…

- When lambda is small (as in the butterfly example), you will often find that your data are overdispersed
  - Your mean $\neq$ variance as is assumed for the Poisson distribution
  - In other words, there is more variation than expected under Poisson
  - This often happens with counts of animals observed at a site

Negative binomial distribution

- Use the negative binomial instead
- In ecology, the NB is mostly used like Poisson, but when you need more dispersion of x

$$\text{Negative Binomial Distribution} = \frac{\Gamma(k+x)}{\Gamma(k)x!}\left(\frac{k}{k+\mu}\right)^{k}\left(\frac{\mu}{k+\mu}\right)^{x}$$

- μ = expected # of counts (mu in R)
- k = dispersion parameter (size in R)
- Mean = μ; variance = $\mu + \mu^2/k$

Negative binomial examples

- Essentially the same as Poisson, but allows for heterogeneity
  - Number of individuals per patch
  - Distributions of parasites on individual hosts
  - Migrating waterfowl over a site
- All these cases are likely to yield clumped/aggregated counts

Continuous distributions

- Normal (or Gaussian)
  - Already covered
- Beta
- Gamma family (gamma, exponential, chi squared)

Normal distribution review

- Symmetric distribution with 2 parameters: mean (location) & scale (SD)
- In R, use dnorm, pnorm, rnorm, qnorm commands

Gamma family of distributions

- Bounded by zero (no negative values)
- Includes gamma, exponential, and chi-squared (latter 2 are special cases of gamma)

Gamma distribution

- We have a 2 parameter model
  - α, β are the parameters to specify
  - Also known as k & 1/θ respectively
  - α (or k) is shape (# of events)
  - β (or 1/θ) is rate (length per event)
- Don't worry about the functional form, just know the properties & when to use it

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

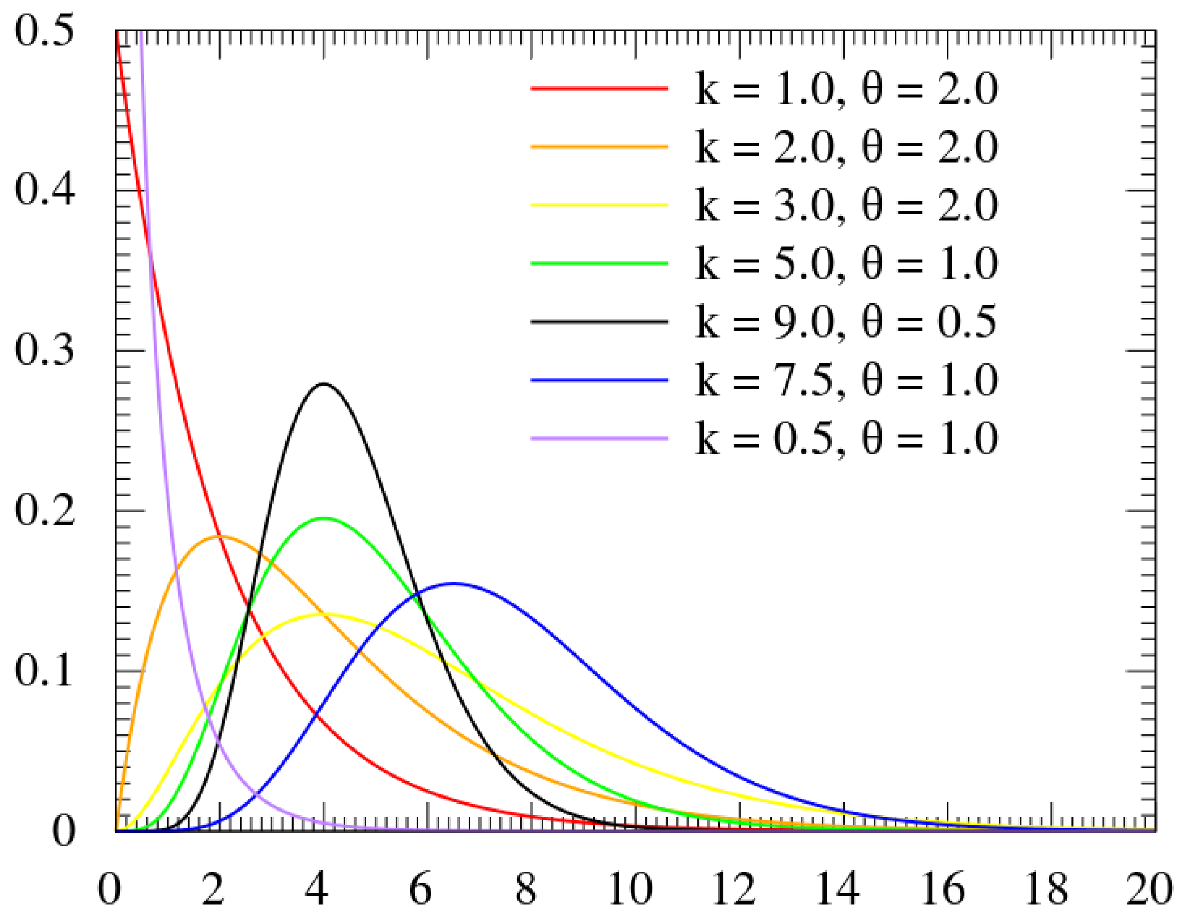
Gamma examples

- May be used to model:
  - Size of insurance claims (e.g. $)
  - Rainfall (e.g. inches)
  - Amount of intact forest (e.g. hectares, acres) surrounding a given site
- Appropriate in cases where you expect overdispersion but it’s a continuous variable so you can’t use NB

Gamma distribution

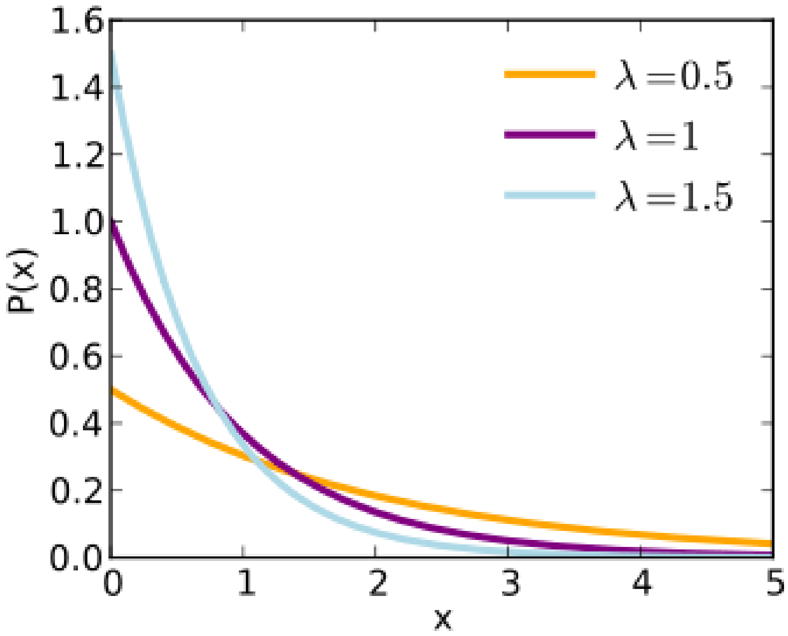
- The skewness is equal to $2/\sqrt{k}$
  - It depends only on the shape parameter (k) and approaches a normal distribution when k is large (i.e. > 10)
- Generally used because it is very flexible in shape and scale
- Useful when data are overdispersed from a normal distribution
- Mean = $\alpha\beta$; variance = $\alpha\beta^2$
- Data you are modeling must be non-negative!

Gamma distribution



Exponential distribution

- Special case of the gamma distribution
- Only 1 parameter: rate λ
- Useful when most probability mass is near zero



Exponential example

- Example: number of miles a car can run on a given battery follows an exponential distribution with an average of 10,000 miles
  - Car owner needs to take a 5,000 mile trip
  - What is the probability he can complete the trip without replacing the battery?

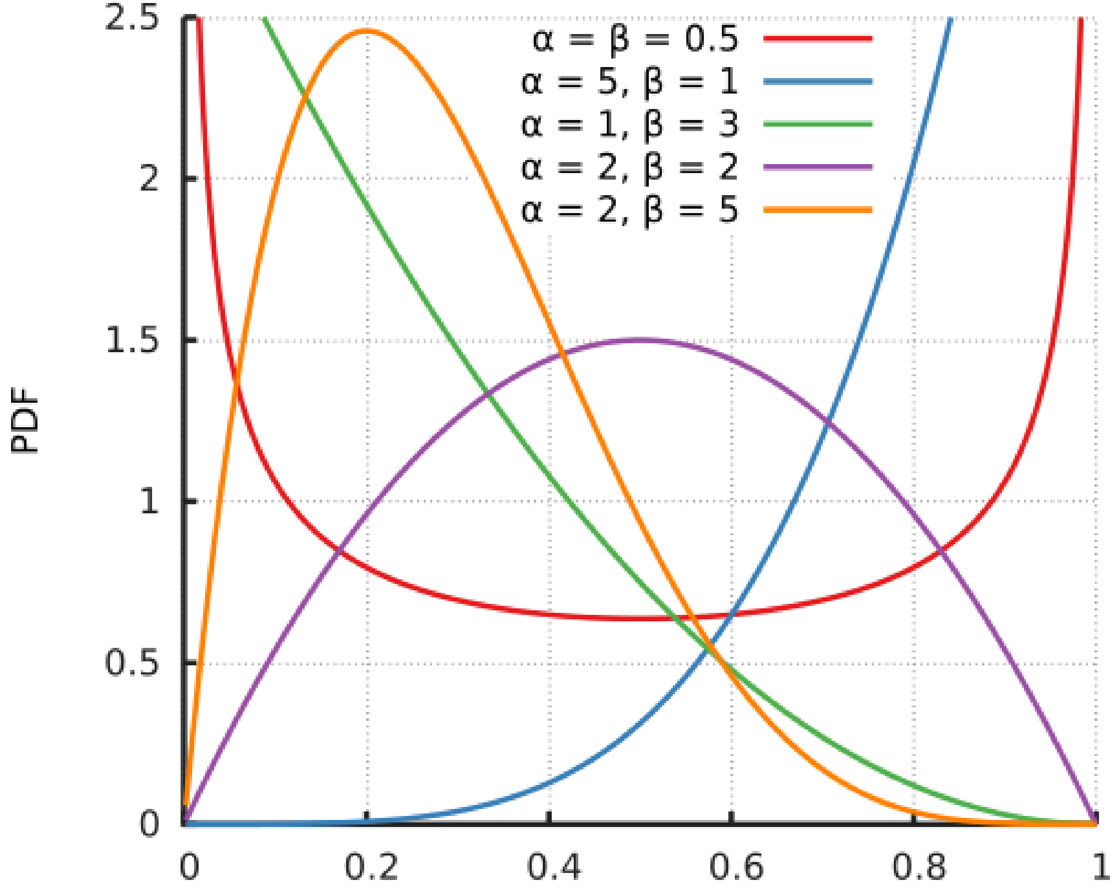
Exponential example

- Let X denote the number of miles that the car can run before its battery wears out
- $P(X > k) = e^{-k/\theta}$ [or $e^{-\lambda t}$ where t is time]
- $P(X > 5000) = e^{-5000/10000} = e^{-0.5}$, which is ~ 0.604

Beta distribution

- Continuous, but constrained on 0,1
- Useful for modeling probabilities or proportions
- Parameterized by 2 shape parameters: α & β
- Has been used for:
  - Allele frequencies
  - Sunshine data
  - Variability of soil properties
  - Proportions of minerals in rocks

Beta distribution

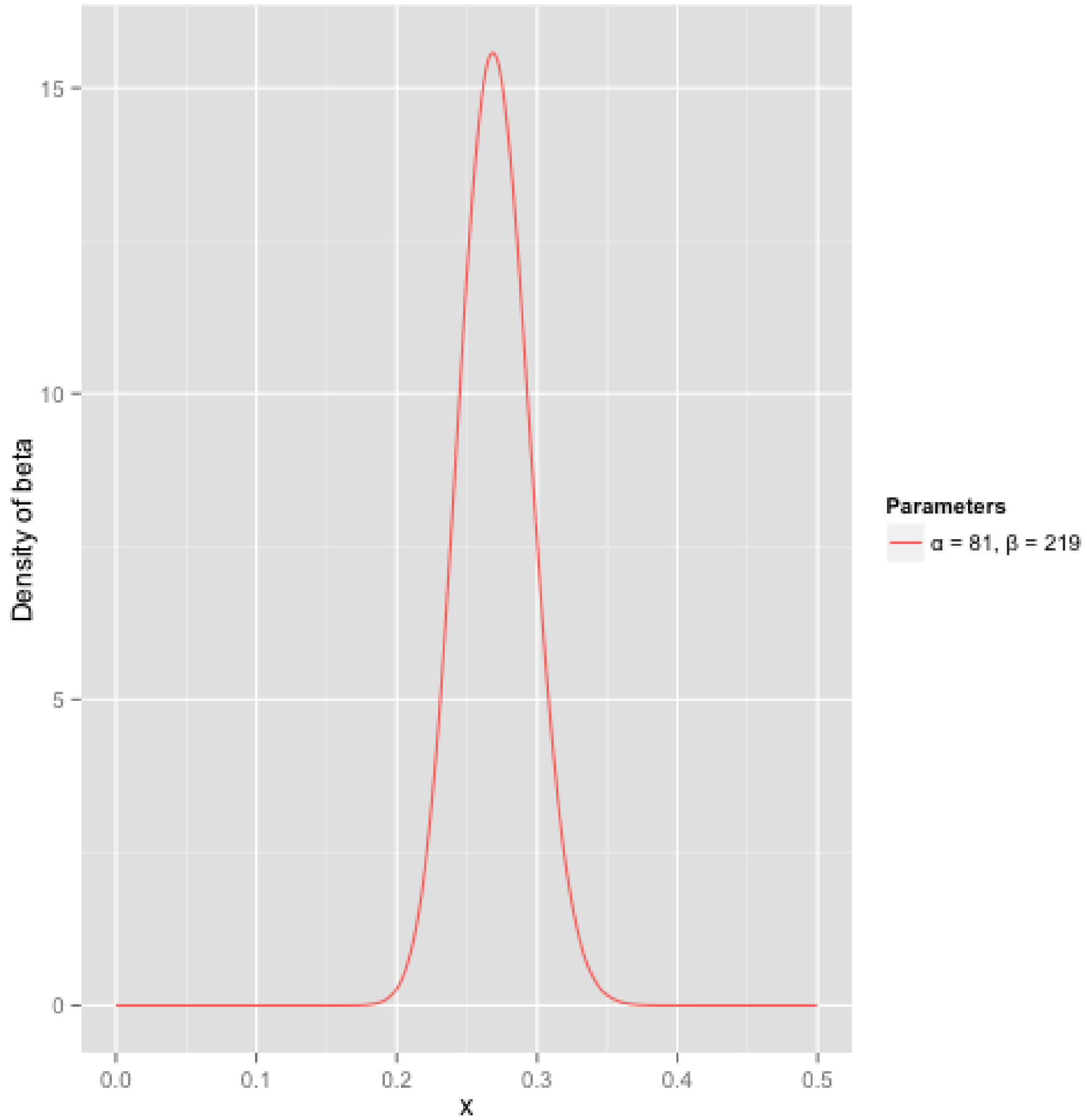


- Mean = α/α+β

Beta example

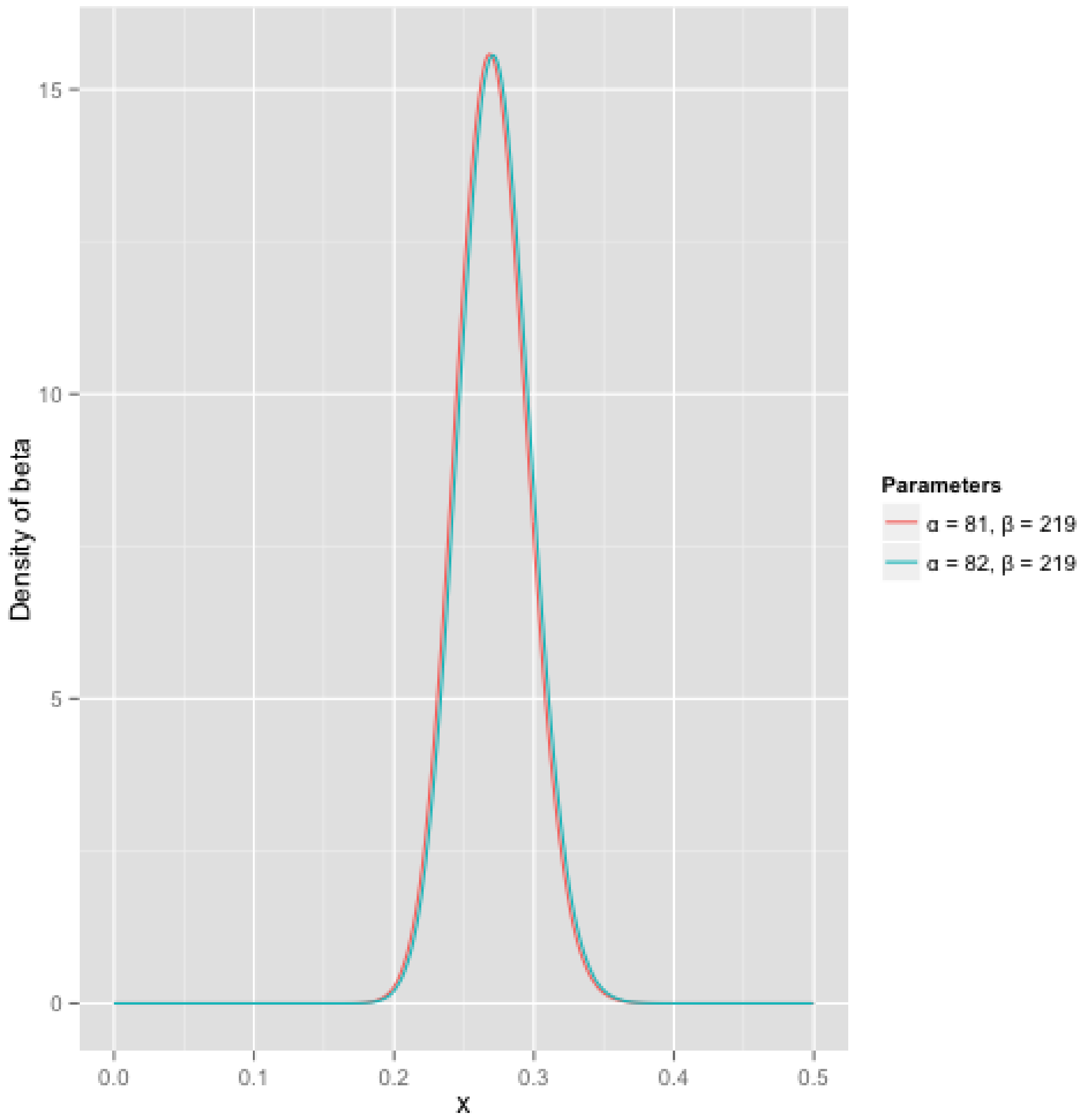
- Let's use baseball statistics to understand the beta distribution
- Batting average: the # of times a player gets a base hit/ # of times he goes up to bat
  - Between zero & one
  - 0.266 is average
- How do we predict a player's season-long batting average?
  - Using the beta distribution to incorporate prior expectations

Beta example

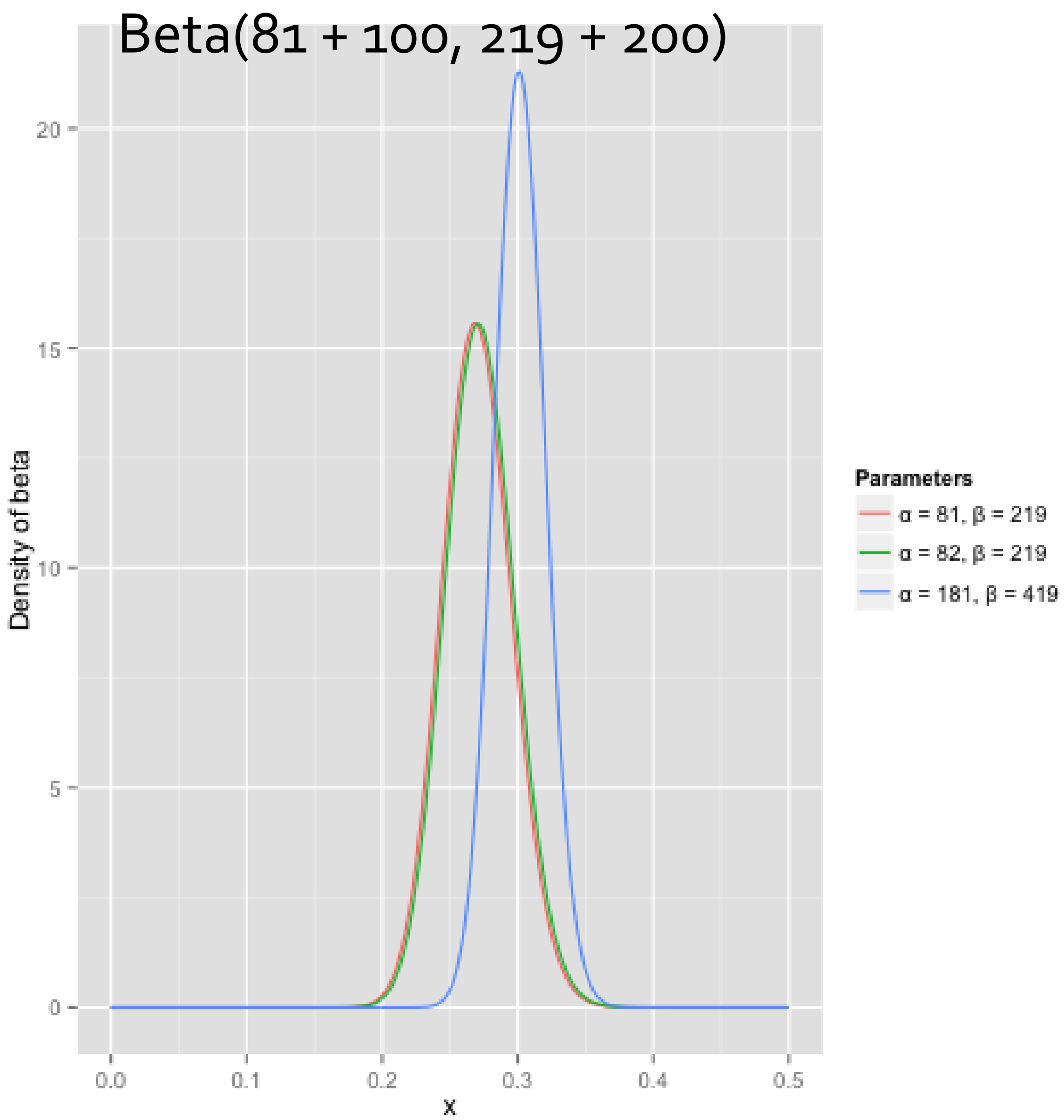


Beta example

Beta(αo + hits, βo + misses)



Beta example



More distributions in R

- Let's go to RStudio to simulate some distributions & learn some tests for normality
- Quiz #2 on Tues (on both probability lectures)
- Assignment #2 due tonight @ midnight