Week 9
Probability and Distributions

Today

- Review homework
- Distributions in ecology
- Distributions in R
- In-class exercise

Why normal data is common

- Central Limit Theorem
  - The distribution of sample means converges to a normal distribution as the sample sizes get larger

Why normal data is common

- The accumulation (sum) of many small random errors drawn from any distribution is normally distributed

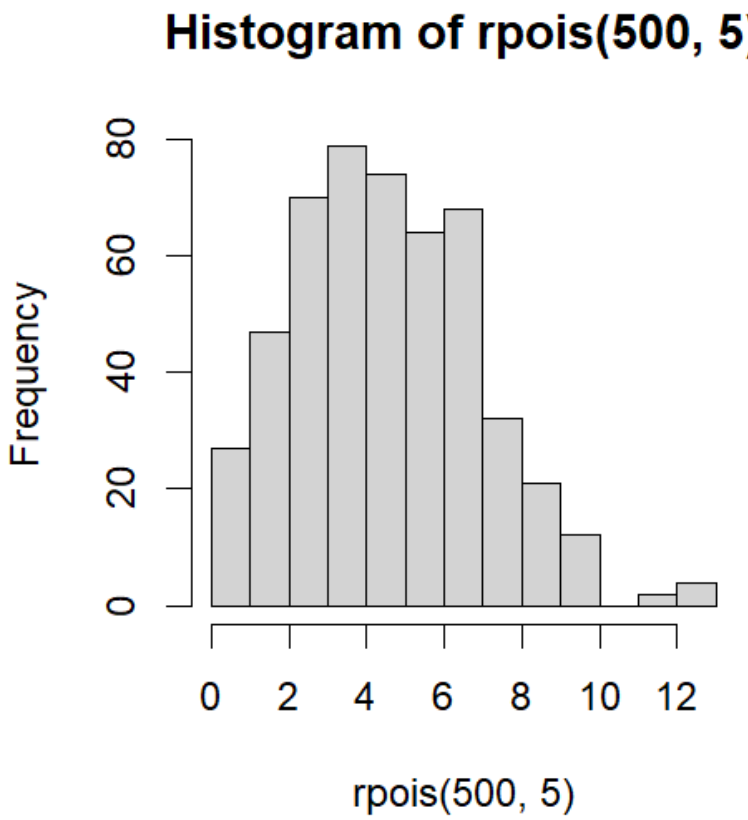
Typical types of non-normal response data in ecological studies

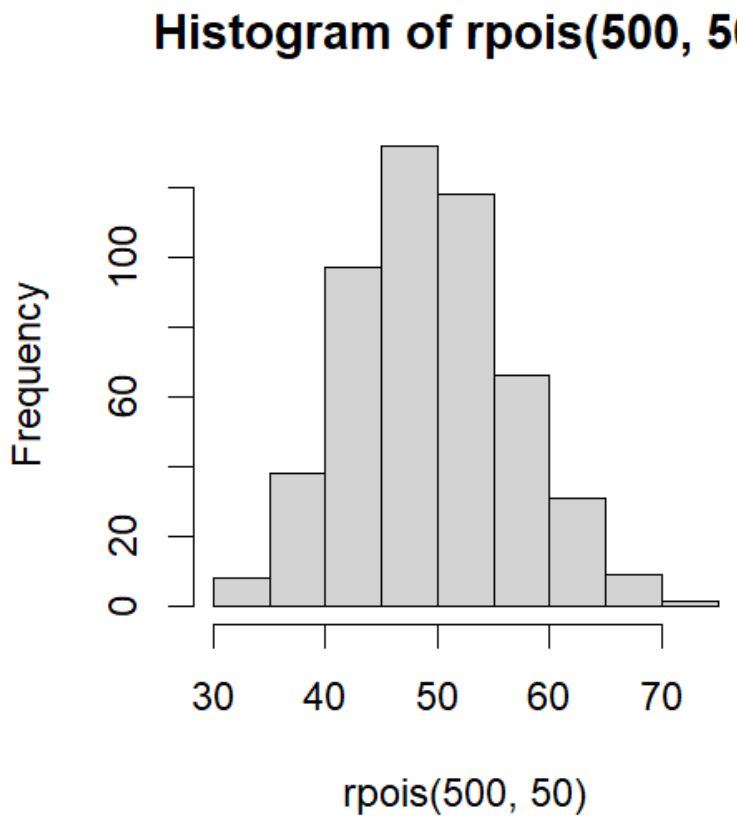
- Counts (quadrats, line transects, etc.)
- Proportions (number of individuals that have some characteristic / total number of individuals examined)
- Categorical (male, female, immature or transitional; benthic, pelagic, and piscivorous morphotypes; canopy, understory, and ground-nesting)

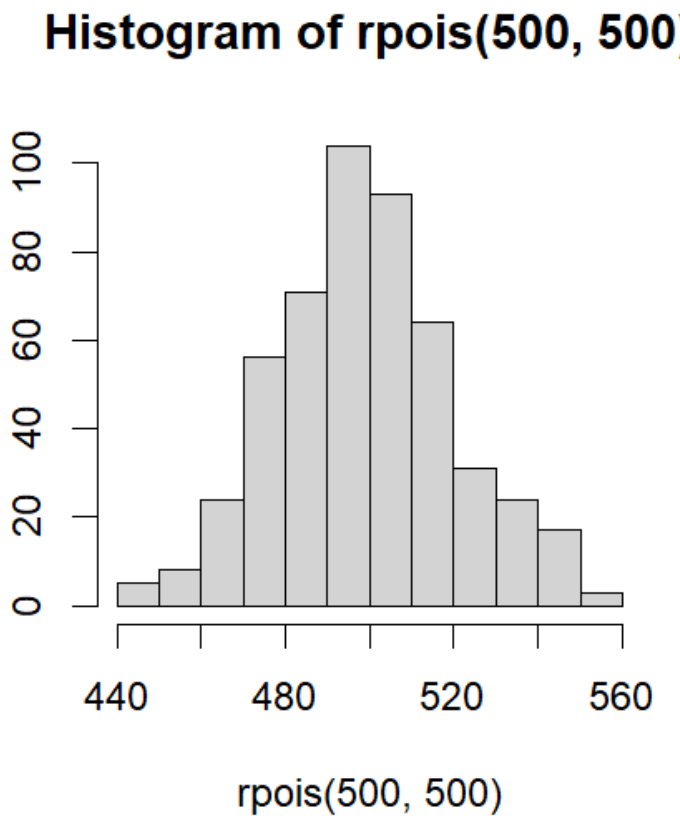
Each of these data types has a typical probability distribution

- Counts of events within a given sampling unit (distance, area, volume, time) are **Poisson** distributed if the rate (density) is constant and they are independent (not clumped in space or time or spread out)
- Discrete probability distribution
  - Integer values from 0 to infinity
- Single parameter: mean = variance = λ
- Clumped data give rise to overdispersion: variance > mean
- **Negative binomial** can be a better option in such cases (two parameters - λ varies in space/time)

Poisson (counts)





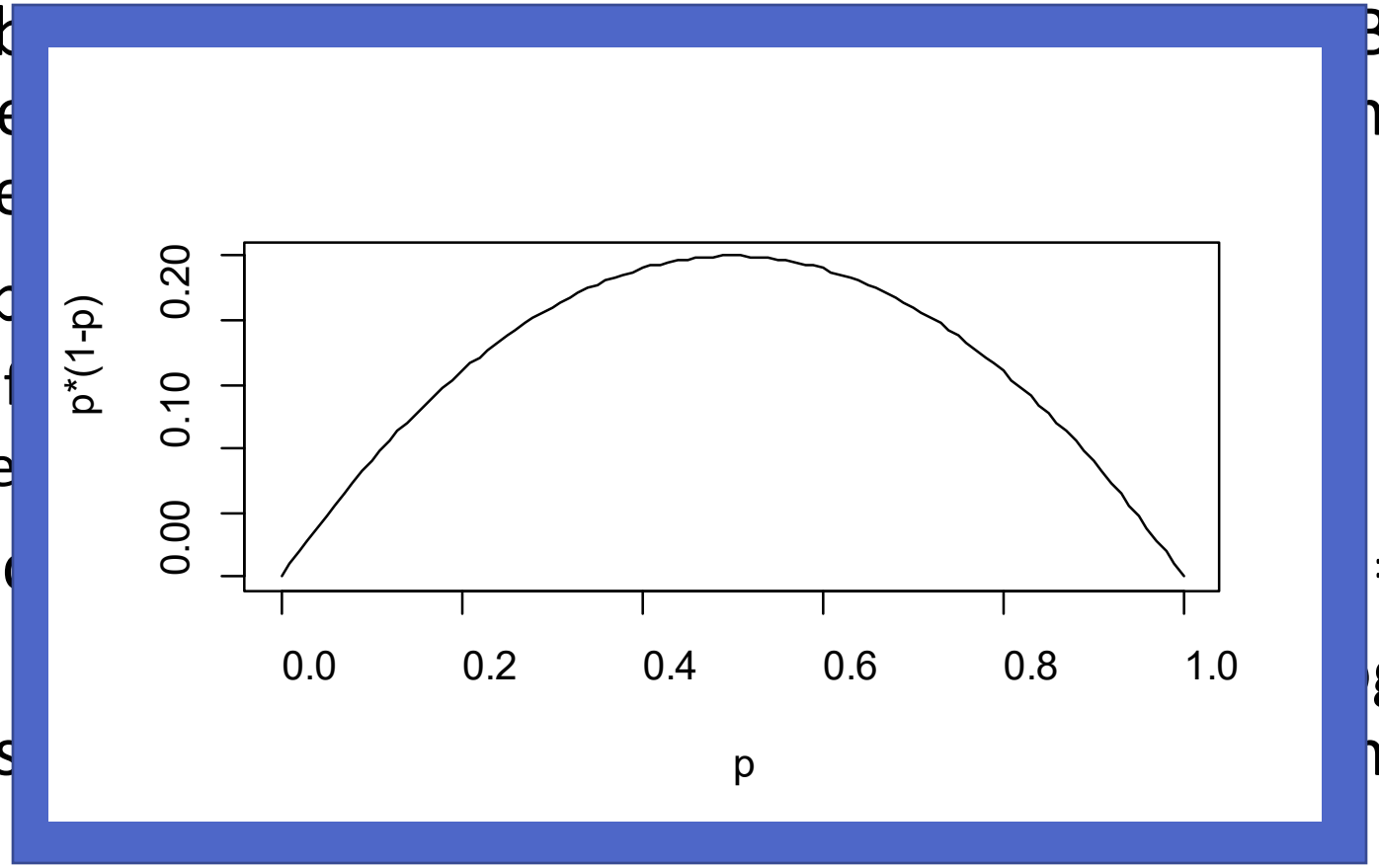


Each of these data types has a typical probability distribution

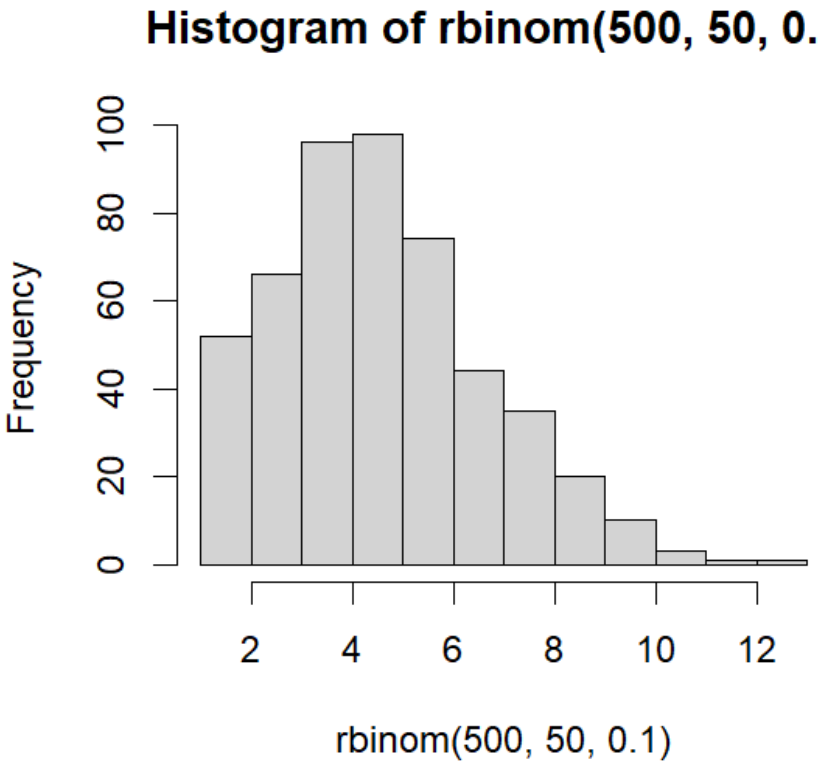
- The number of "successes" in a set of *n* independent "Bernoulli trials" or "coin flips" – i.e. observations are binary (yes/no, success/failure, male/female) follow the **binomial** distribution
- Discrete probability distribution
  - Values from 0 to *n*
  - Parameters *n* and *p*, the probability of "success"
- Variance changes with the mean: mean = $n*p$ variance = $n*p(1-p)$
- From the equation (or from considering the problem logically) can you guess the value of p at which the variance is maximized?

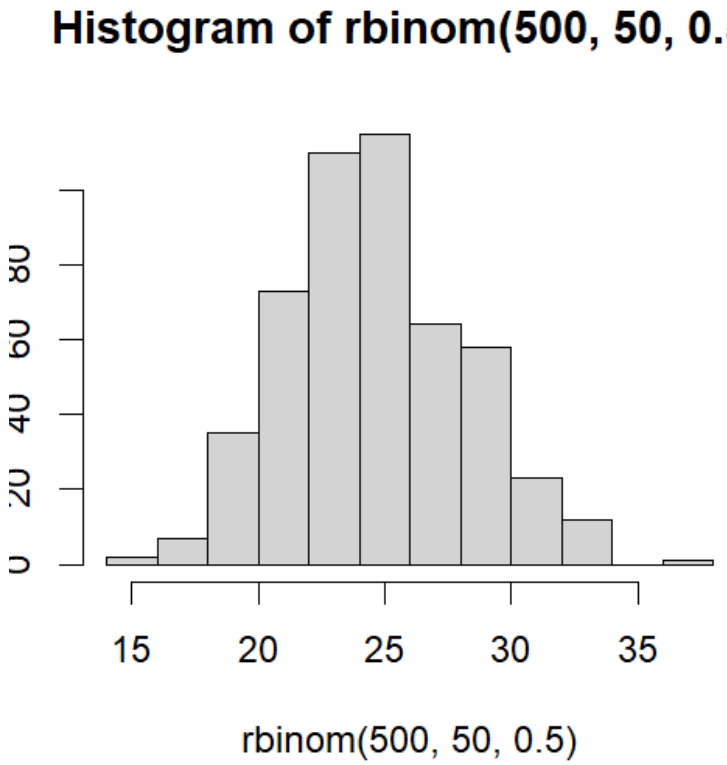
Each of these data types has a typical probability distribution

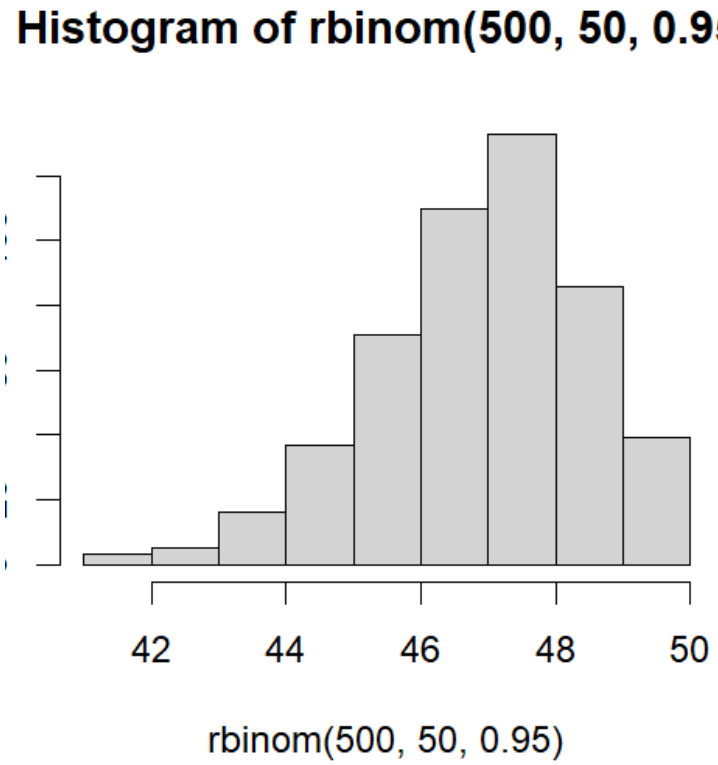
- The numb… Bernoulli trials" – i.e. obse… male/female) follow the…
- Discrete p…
  - Values …
  - Parame…
- Variance … = $n*p(1-p)$
- From the … gically) can you guess… nized?



Binomial (number of successes/failures)



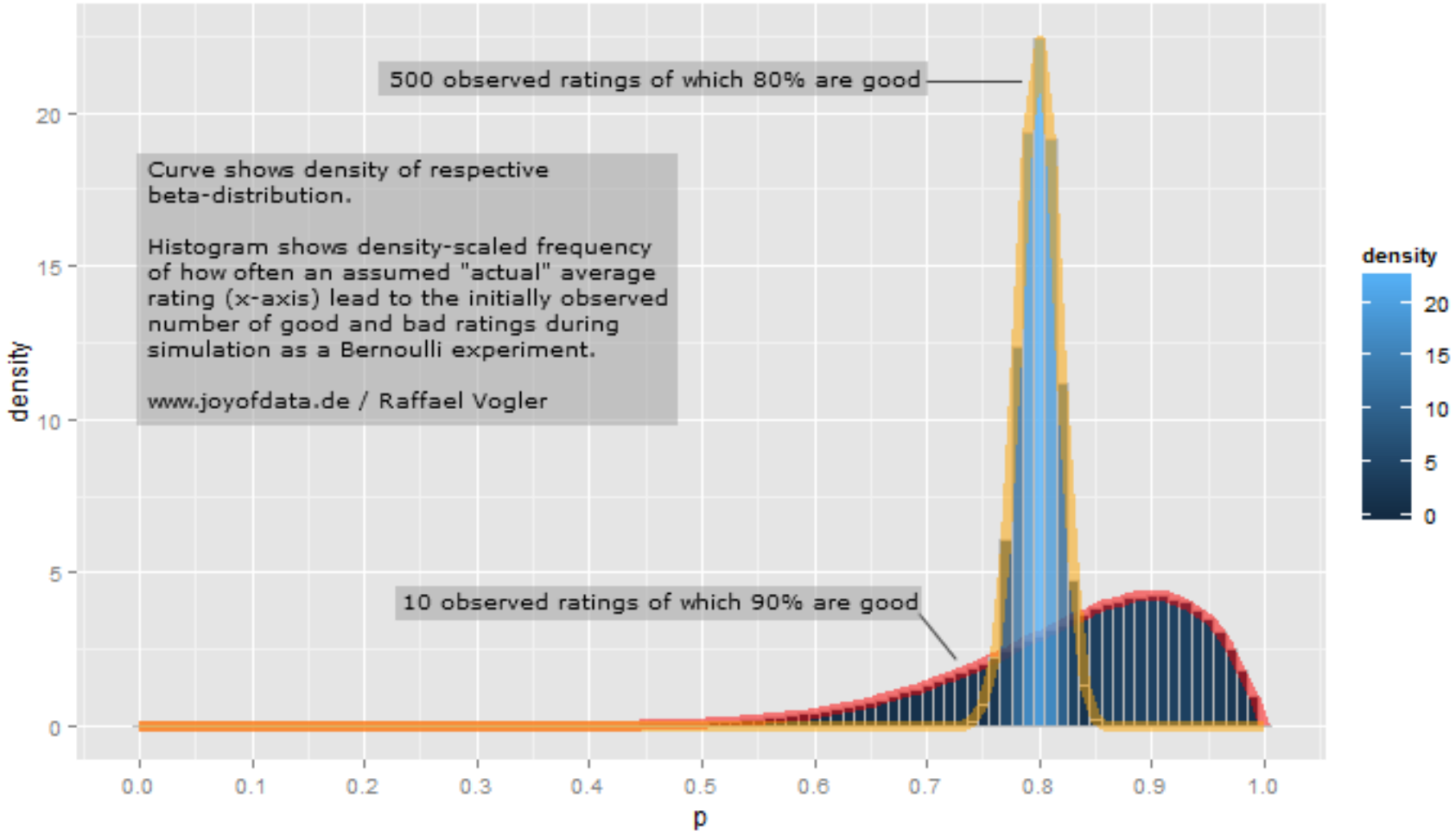




Each of these data types has a typical probability distribution

- The distribution *p* can be defined based on the results of a set of *n* independent “Bernoulli trials” using the **beta** distribution
- Continuous probability distribution
  - Values from 0 to 1
  - Parameters *alpha* and *beta*
- Distribution of p is a beta with alpha-1 “successes” and beta-1 “failures”

Each of these data types has a typical probability distribution



Exponential family

- Normal (Gaussian), Poisson, Binomial, Gamma, Inverse Gaussian
- Can be expressed as a function of two parameters:
  - Θ, the “canonical parameter” or location
  - Φ, the “dispersion parameter” or scale
- And three functions: a, b, and c

$$f(y|\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

Distributions in R

- rnorm() – draw random variables from the specified distribution
- dnorm() – the probability density at any point
- pnorm() – tail probabilities to the left or right of the specified point
- Similar r-, d-, and p- syntax for other distributions: e.g., rpois, runif, etc.

Distributions in R

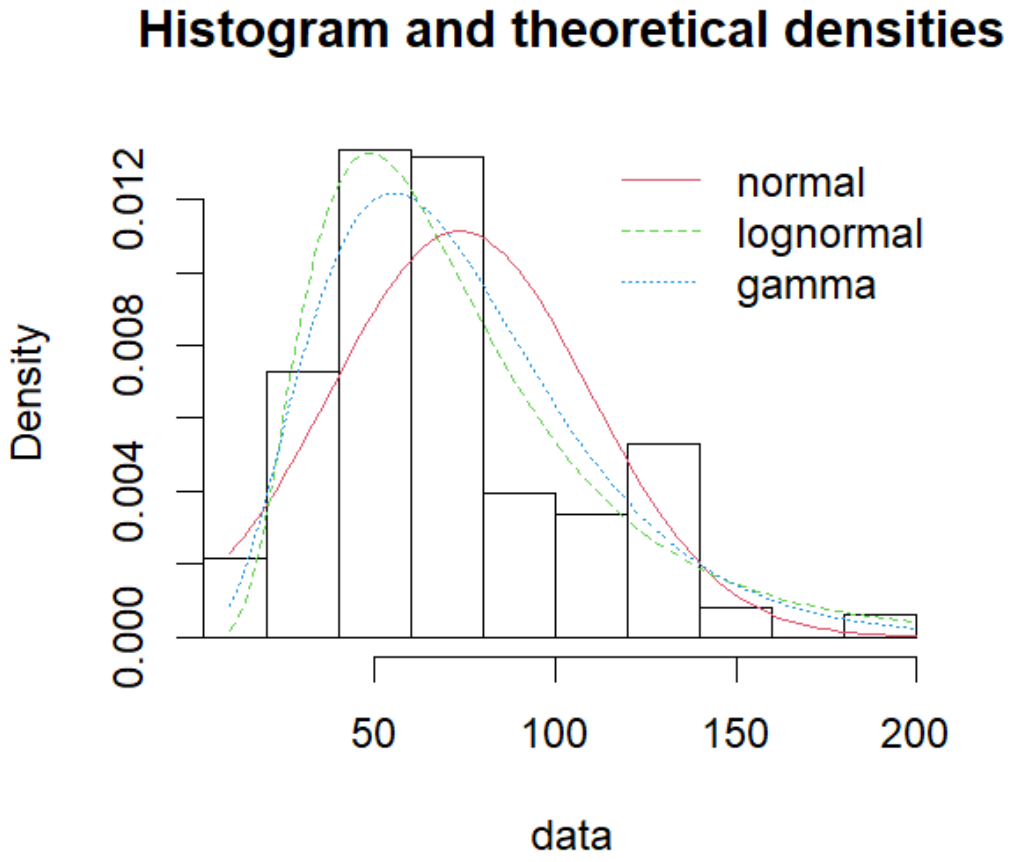
- How can our results be replicable if they involve generating random observations?
- set.seed()
- Sets a starting number. Same results every time this number is the same.

Using random numbers in R to select a random subset of observations

- Create a new column in your df
- Fill it with random numbers drawn from a continuous distribution (so no ties)
- Sort by this new column
- Select the first X rows

Fitting distributions and distribution tests

- MASS and fitdistrplus packages
- denscomp() plot function



Exercise: Is this true?

- The accumulation (sum) of many small random errors from any distribution is normally distributed

- What if the process is multiplicative rather than additive – i.e., the product of many small random errors…

- Use fitdistrplus to compare distributions

Exercise: Is Don Corleone cheating?

- Don Corleone challenges you to flip a coin three times. Each time the loser owes the winner a favor

- After four heads, you owe him four favors. And you're wondering if the game is rigged

- What is the probability that the probability of heads is > 0.5