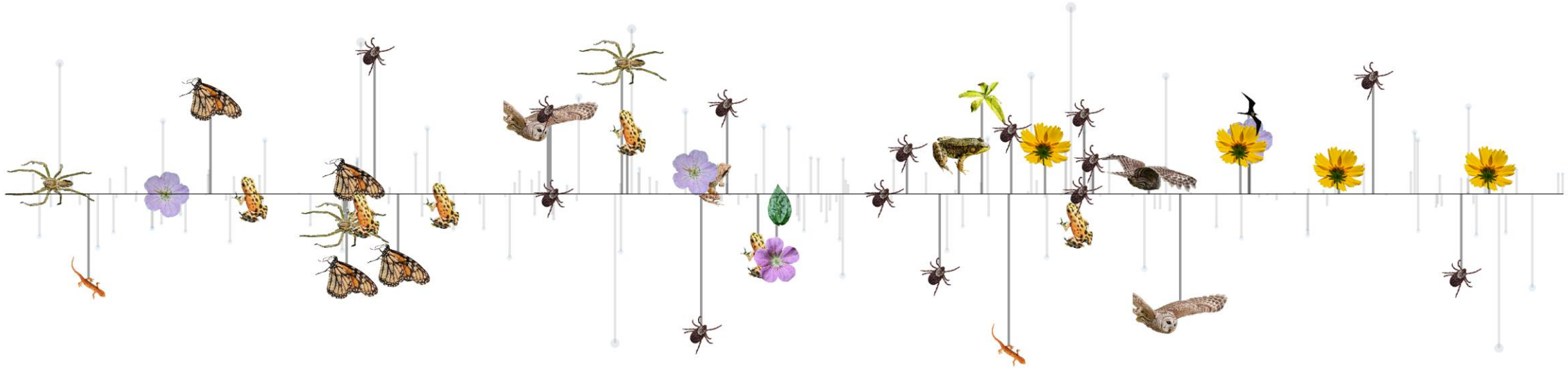
Intro to Quantitative Ecology
UMass Amherst – Michael France Nelson

Deck 4: Numerical Data Exploration



Announcements

- Chapter 4 has a lot of important information.
  - We will move through it quickly, but we are going to revisit and re-use concepts throughout the rest of the course.

Numerical Data Exploration: Model Thinking

We may not realize it, but we are engaging in model building when we summarize data.

Some of our assumptions include:

- Numerical and graphical summaries and exploration are a valid way to characterize data. (they usually are)
- We think (or hope) our data are representative.
- We think that *statistics* like mean or standard deviation tell us something *meaningful* about our data.

Numerical Data Exploration: Sample Statistics

What are two general quantities to summarize a collection of numbers?

- Central tendency
- Dispersion

Why do we call these *statistics*?

Population

- A large collection of sampling units
- We usually can't observe the entire population
- Properties of the population are called **population parameters**

Sample

- A subset of the population
- We can observe the entire sample
- Properties of the sample are called **sample statistics.**

Numerical Data Exploration

Some tools and statistics:

- 5-number summary
- central tendency: mean, median, mode
- spread/dispersion: standard deviation, range
- min, max
- skew. We don't often formally quantify this, but we frequently consider it *graphically*.
- tests for normality, like shapiro.test() in R

Center and Spread

Center	Spread
• A measure of the characteristic value of a collection of numbers. • What number are we most likely to observe if we choose one randomly? • Mean • Median • Mode	• A measure of the dispersion. • How variable are the values in a collection of numbers. • Standard deviation • Range (minimum and maximum) • Interquartile range

Distributions

What are key features of the Normal Distribution?

What is the Uniform Distribution?

Normal	Uniform
• Hump-shaped • Symmetrical • Two parameters: mean and standard deviation • Most values are near the mean • The standard deviation determines the width of the normal	• Flat-shaped • Symmetrical • Two parameters: min and max • Values are evenly distributed: no value is more or less likely than any other

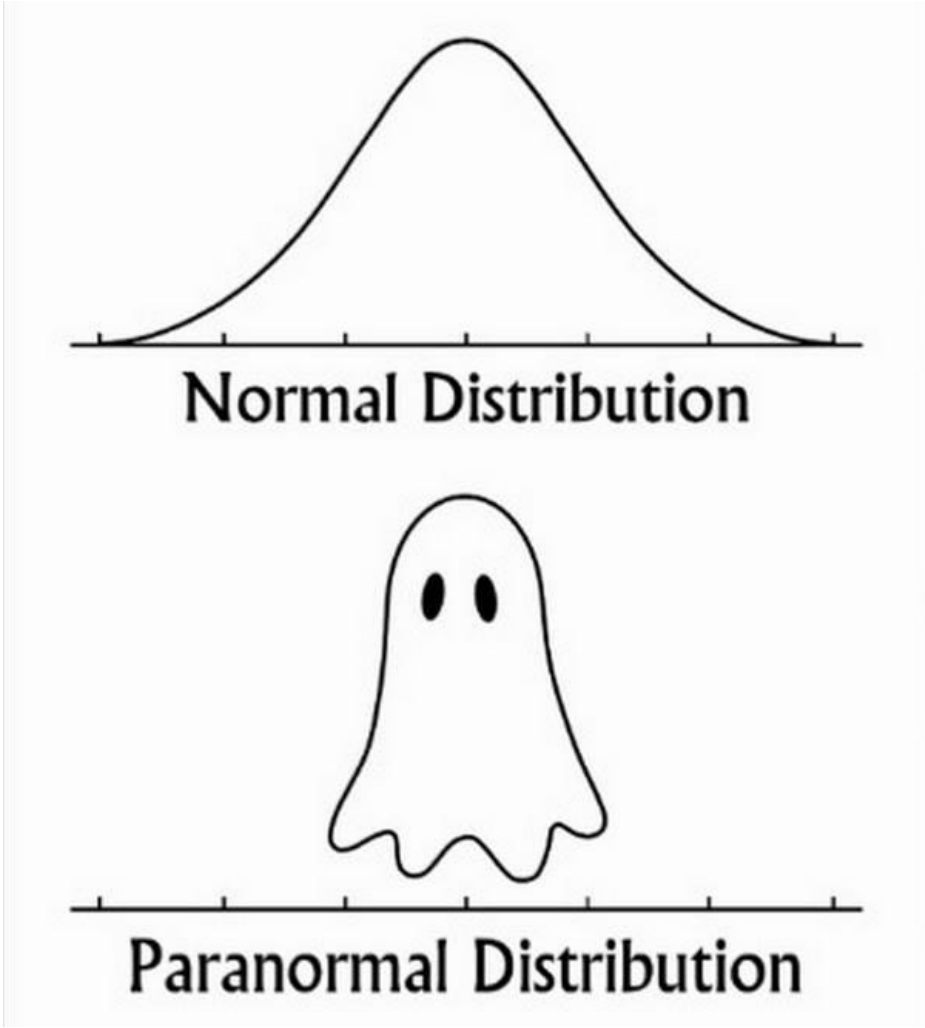
Distributions

Your book greatly simplifies the concept of distributions…

- Gardener calls the Normal distribution *the parametric distribution*.

Look at table 4.11:

- Do you really think there are only two distributions?
- Hint: there are hundreds of named distributions…



The World of Distributions

Ask yourself:

- Does the Normal work for binary (true/false, presence/absence) outcomes?
- Does the Normal work for categorical data?
- Does the Normal work for count data?
- There are hundreds of *parametric distributions* that can model these scenarios (and others).

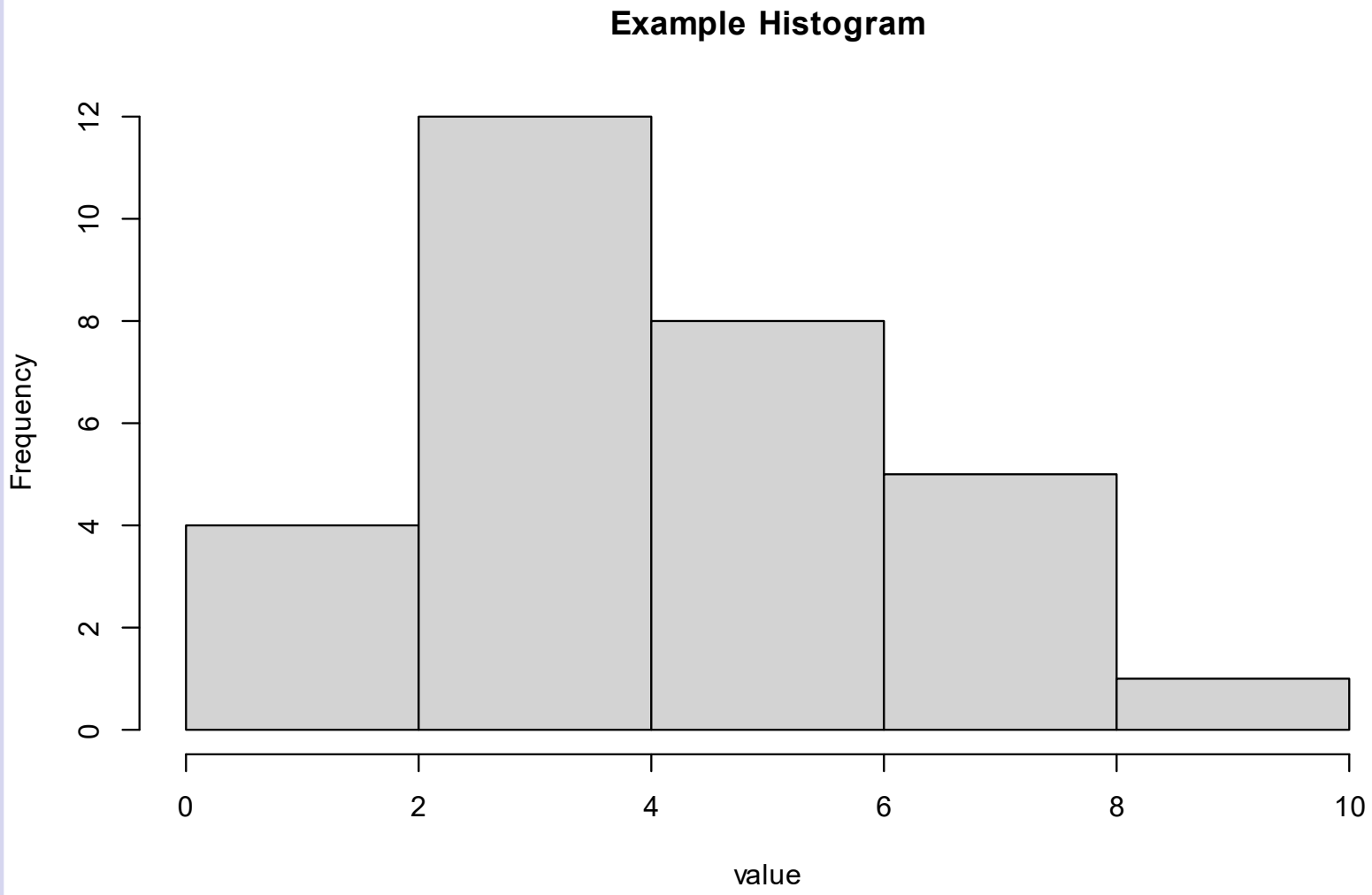
Some Other Important Distributions

- Bernoulli and binomial:
  - These model the number of *successes*.
  - Think of flipping a coin one or more times and counting the number of *heads* or the number of plots in which a species is present or absent.
- Poisson: modeling count data
- Exponential and geometric: modeling skewed data in which small measurements are most common.
- T-distribution: a small-sample version of the Normal.
- Chi Square and F: sums and quotients of multiple normal distributions. Used for lots of statistical tests.

Histograms

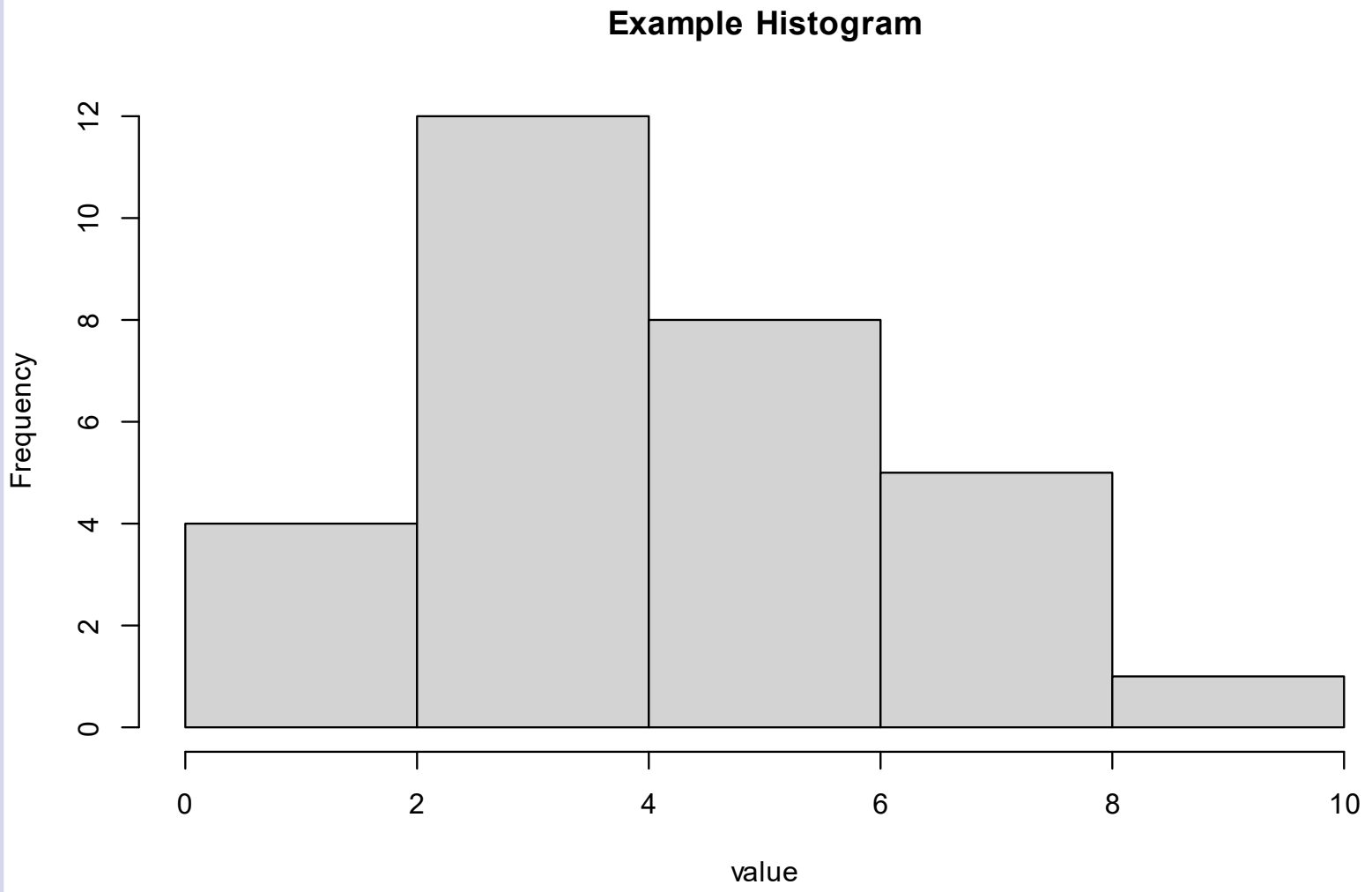
Graphical Exploration: Preview – The Histogram

- Histograms are very powerful, but often mis-interpreted
- Histograms **do not** plot raw data, they plot a **summary**



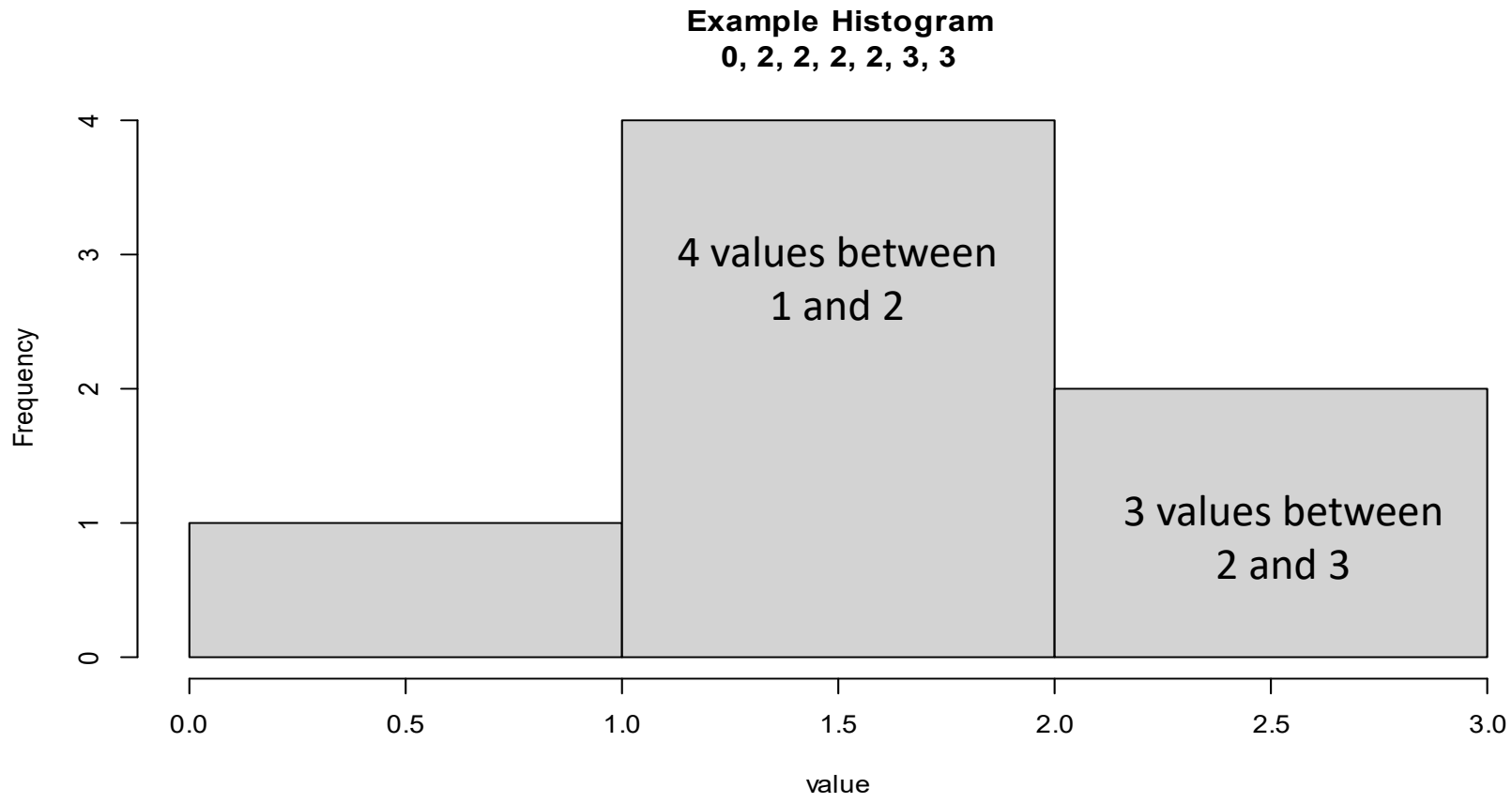
Graphical Exploration: Preview – The Histogram

- Histograms help us understand the **distribution** of a collection of numbers.
  - They are similar to the plot of a **distribution function** (we'll learn about these later)



How To Read A Histogram

- Bins heights (y axis) are counts
- Bin widths (x axis) are data value ranges



Random Numbers

Random Numbers in R

What are random numbers?

Random numbers generated by a computer are actually pseudorandom.

- Can our computers really generate randomness?

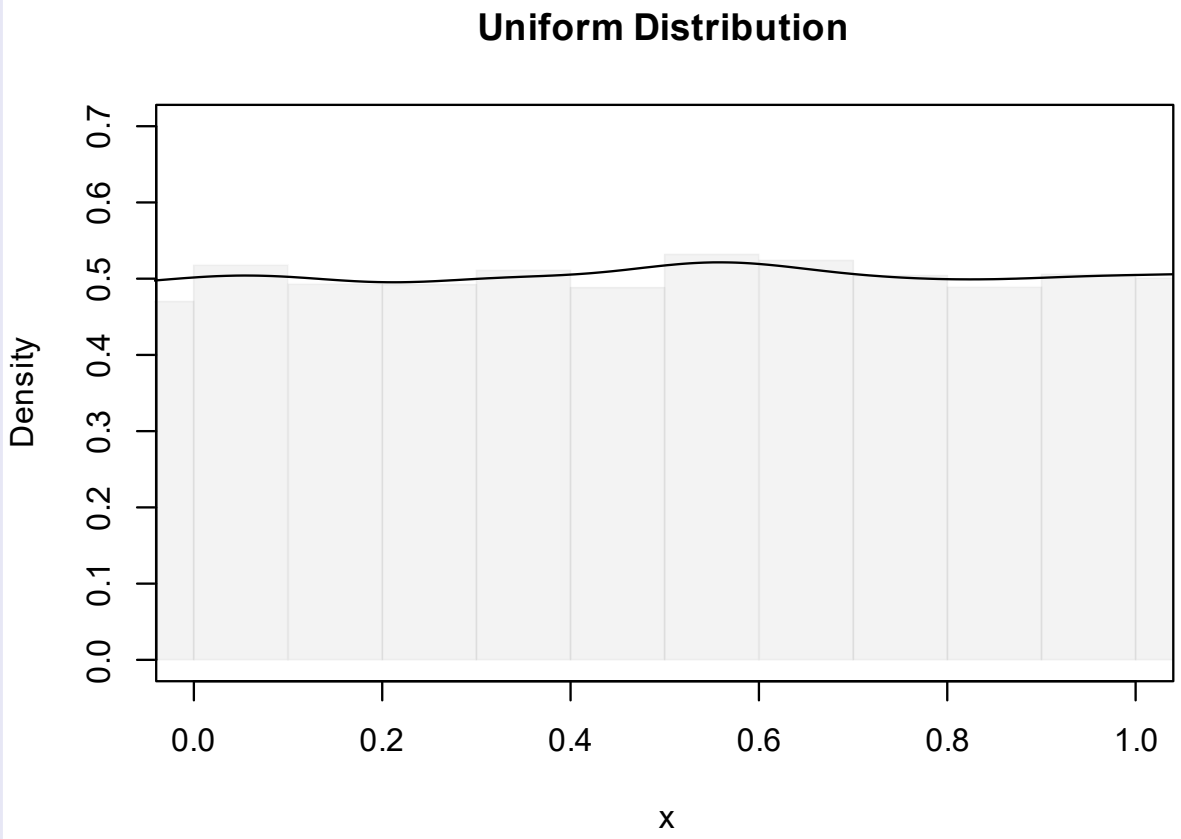
Random Numbers in R

What is pseudorandom?

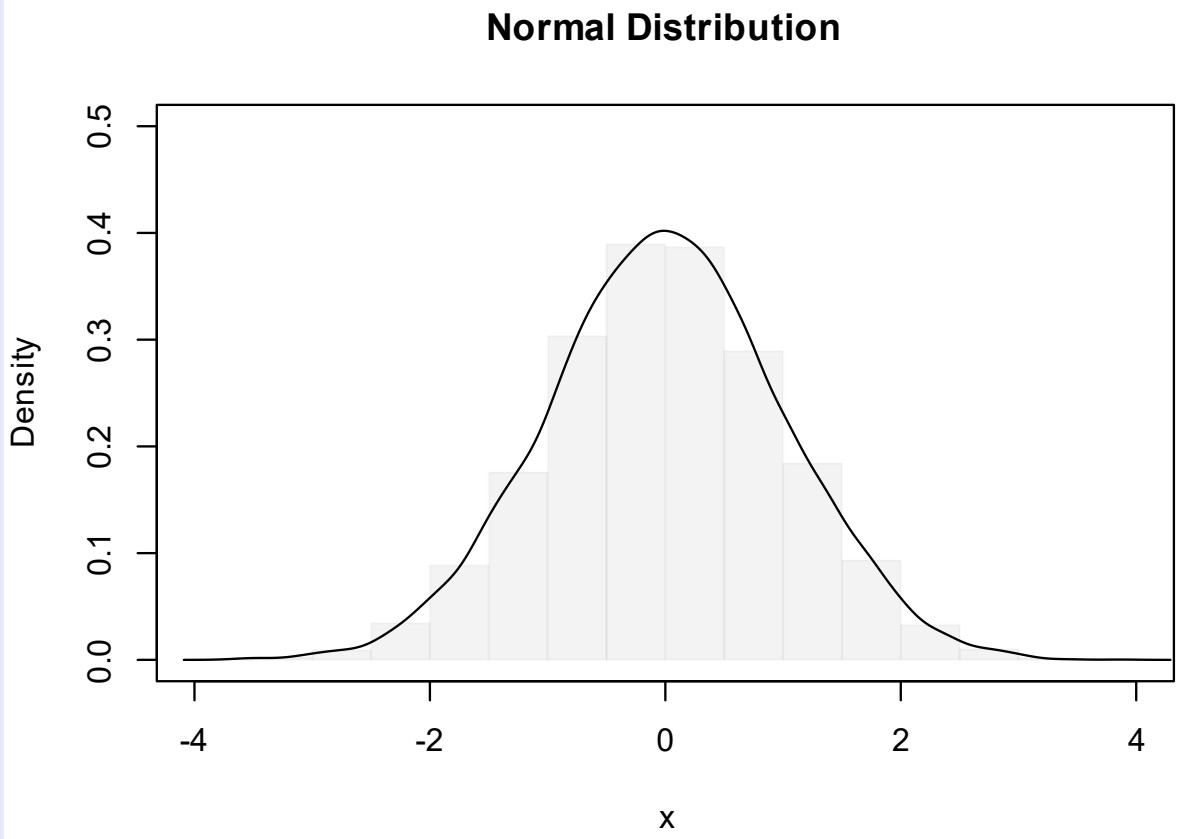
- Sequences of numbers that match the statistical properties of randomness.
- Generated by numerical algorithms, initialized using seed numbers.
  - set.seed() in R

Distributions: Preview – 2 Important Distributions

Uniform Distribution: All numbers are equally likely to occur



Normal Distribution: Numbers close to the mean are more likely to occur



Sampling From Distributions: The r Functions

Normal Distribution	Uniform Distribution

`rnorm(`
`  # How many numbers to generate`
`  n = 100,`
`  # The center`
`  mean = 0,`
`  # The dispersion`
`  sd = 1`
`)`
	`runif(`
	`  # How many numbers to generate`
	`  n = 100,`
	`  # The minimum possible value,`
	`  min = 0,`
	`  # The maximum possible value`
	`  max = 1`
	`)`

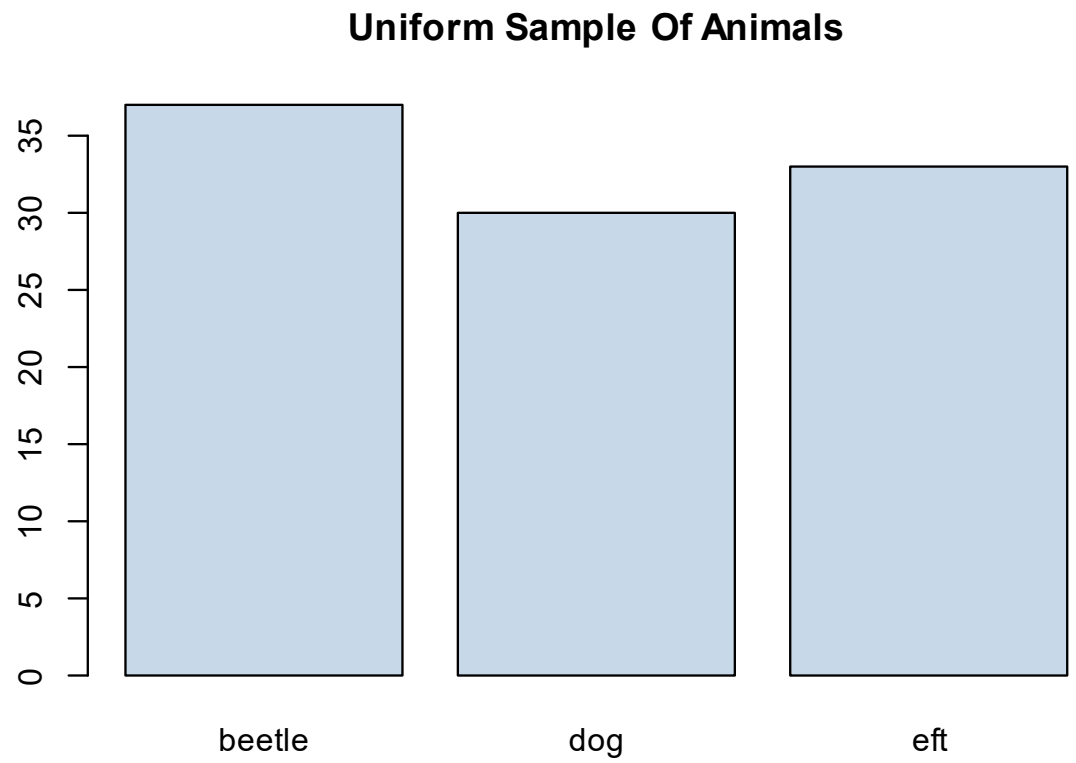
Sampling From a Vector: sample()

Syntax Example

```
# Set the random seed
set.seed(371521)
set.seed(1)
animals = c("eft", "beetle", "dog")

# Randomly sample 50 animals
animals_s = sample(
  x = animals, size = 50, replace = T)
head(animals_s)
barplot(
  table(animals_s),
  main = "Uniform Sample Of Animals",
  col = adjustcolor("steelblue", 0.3))
```

Barplot



Questions for me?

- General questions?
- R questions?
- R demos?

Random Number Generation

- Take 5 minutes to read through the instructions.
- Submit your report at the end of class.
- Be sure to include your names in the report.
- Feel free to shuffle groups!