Week 4: Data Management

Session 2

iClicker questions

Which of the following are *not* one of the statistical concepts we have discussed in the last few lectures?

A statistics and parameters

B populations

C samples

D variance and standard deviation

E center and spread

Announcements

No class on Tuesday!

Chapter 4 has a lot of material. I'm moving the due date for the pre-class reading assignments to Wednesday of next week (by midnight).

Distributions

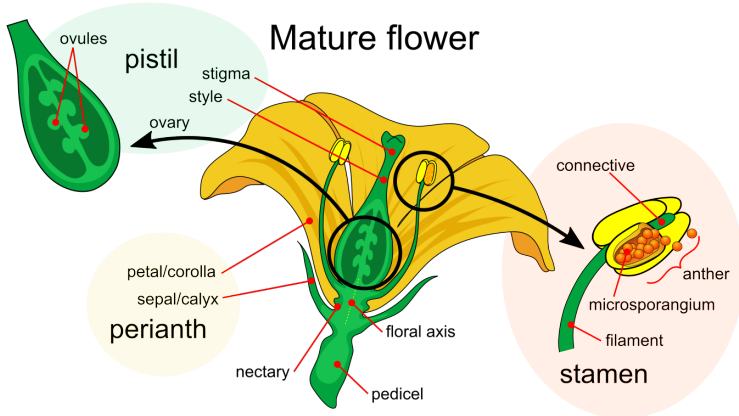
Your Gardener book mentions the Normal distribution in chapter 4, but there is no prior discussion about what *distribution* means!

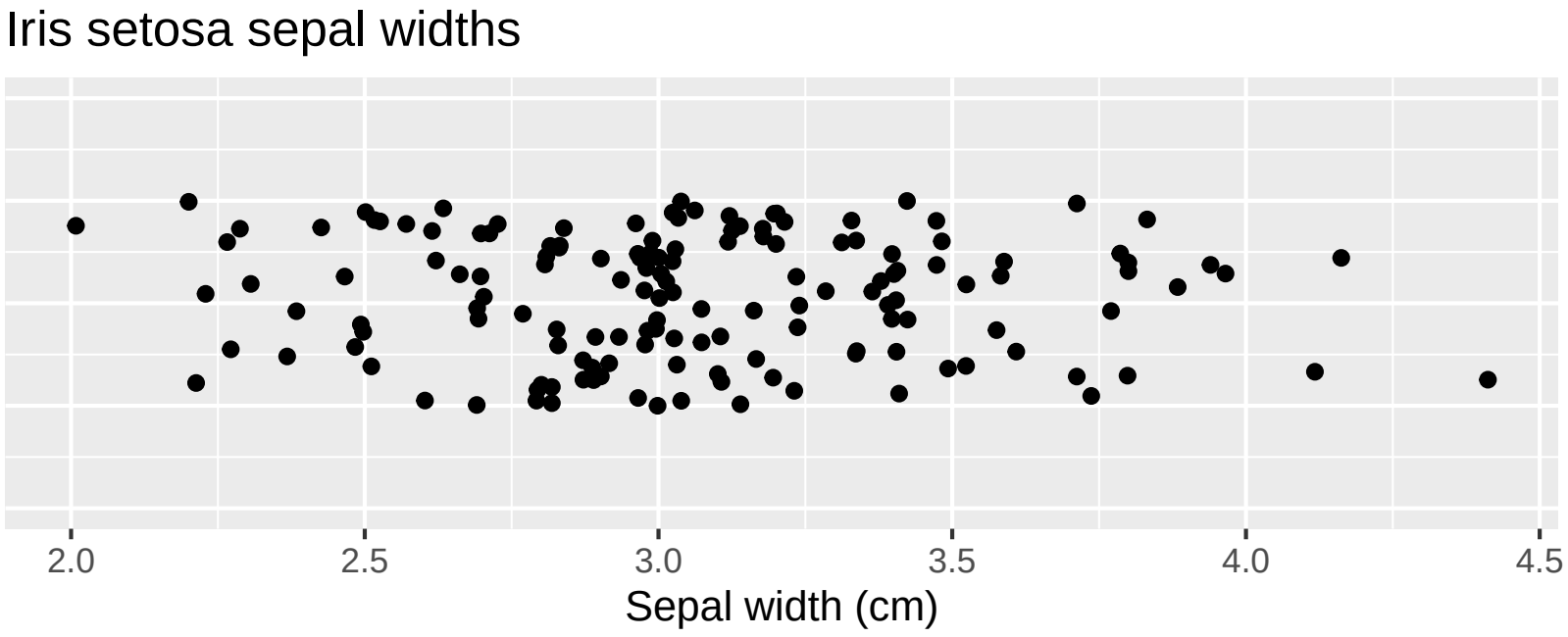
What is a distribution?

We're going to build an intuitive concept of a distribution from the ground up.

Distributions

How variable are the sepals of in a wild population of *Iris setosa*?



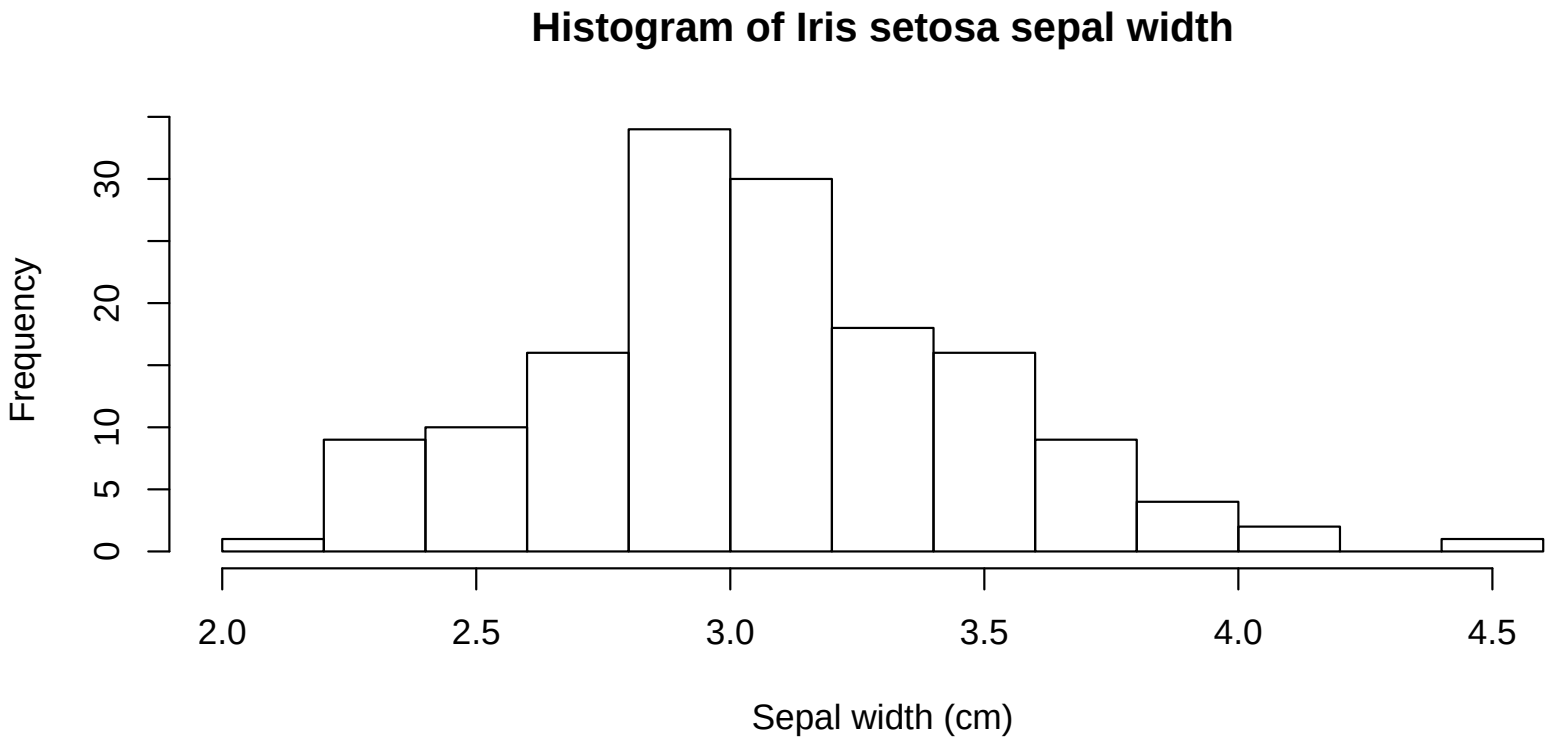


Distributions

What is the **center** of the *Iris setosa* sepal width data?

What is the **dispersion** or **spread** of the *Iris setosa* sepal width data?

A *histogram* plot displays the relative frequencies of size *bins*.



Distributions

A **distribution** describes **variability** in a collection of data points.

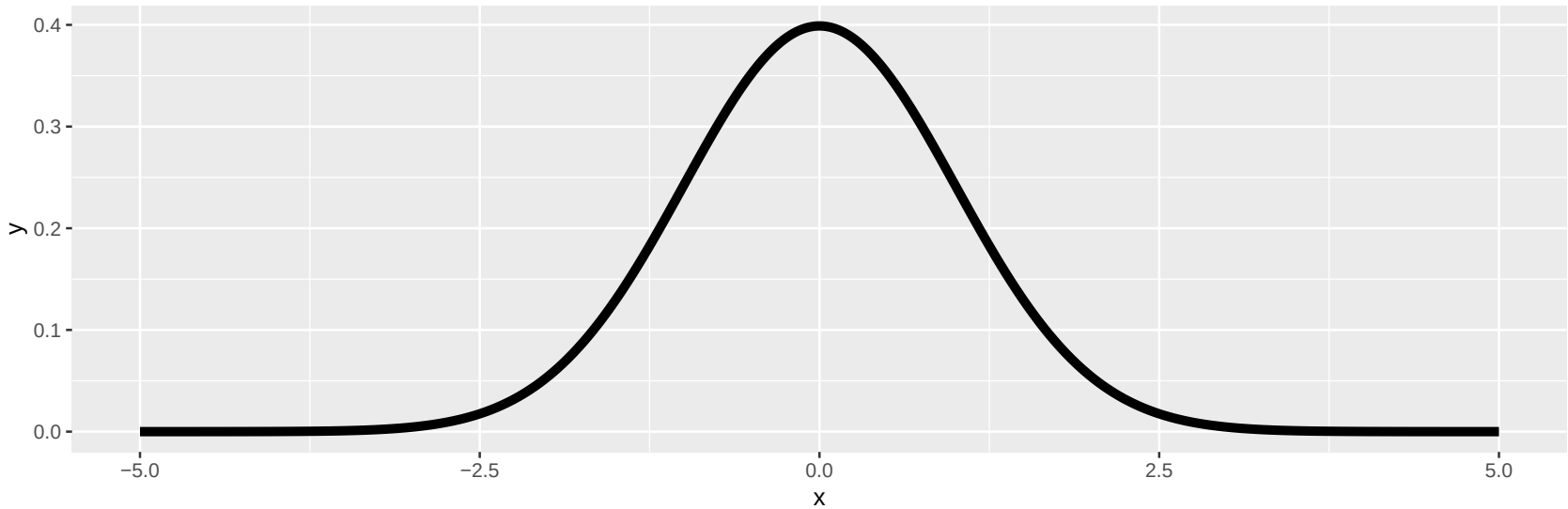
There are many kinds of distributions.

The Normal Distribution

You've probably heard of the **Normal** distribution.

It describes a collection of data that are centered around a most likely value, the *mean*. The likelihood of values away from the center decreases *symmetrically* with the magnitude of the distance from the mean.

It has a familiar bell shape:

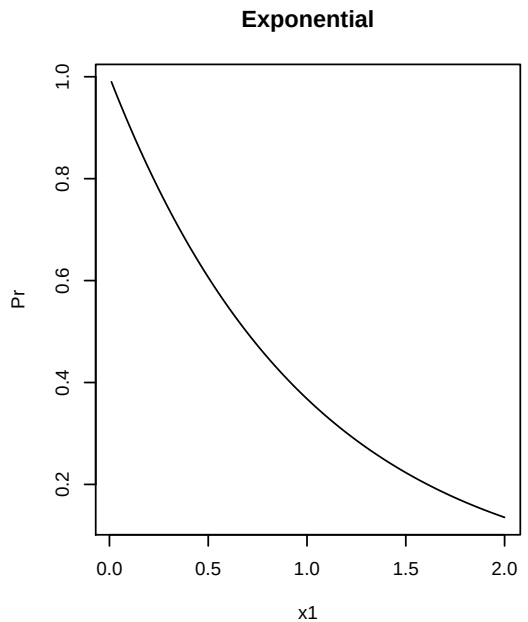


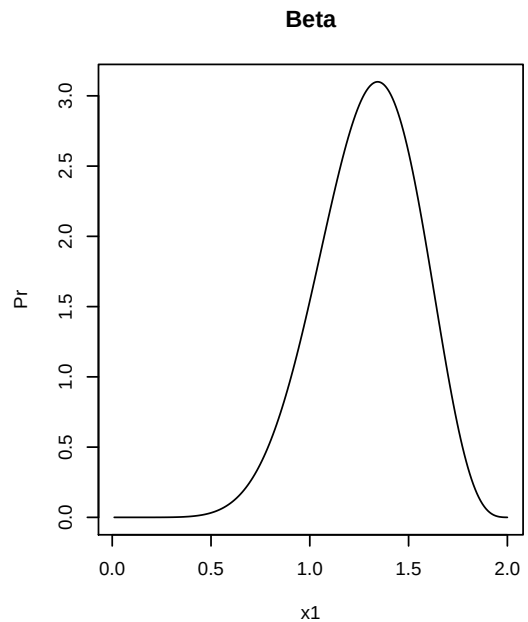
Distributions

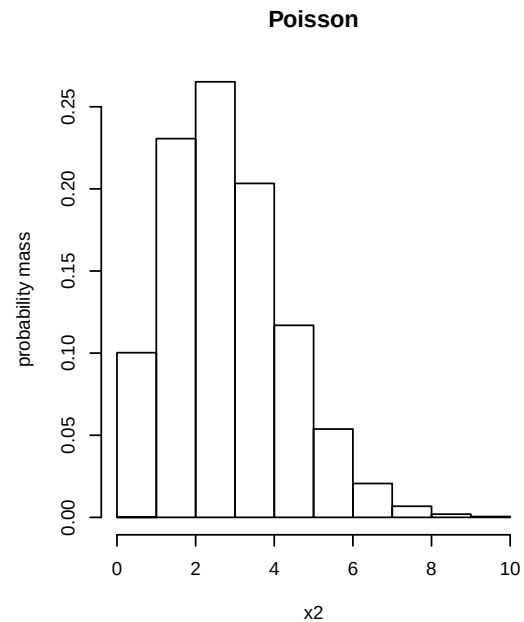
Different distributions describe different kinds of variability.

We'll encounter many types throughout the course.

Here are 3 of the most famous:







Center and Spread

What are some statistics we use to measure center?

How could we describe spread?

Center and Spread in R

These functions can help:

```
mean()
range()
median()
min()
max()
sd()
summary()
```

Group Activity: Salamanders

We have data for salamanders collected at 4 locations in the Pioneer Valley.

The data include information about:

- Snout to vent length
- Total length
- Sex
- Observation season

You can find the data file and detailed instructions on Moodle.

Today you'll do some exploratory data analysis. Load the data into R and use some of the functions you know to examine the data.