

Analysis of Environmental Data

Data Exploration, Associations, and Functions

Michael France Nelson

Eco 602 – University of Massachusetts, Amherst
Michael France Nelson

Announcements

- Updated Deck 3 slides: re-download!
- DataCamp: Intermediate R
 - Optional, but may be useful as a supplement to the course/lab materials
- In-Class Model Thinking
 - Graded, except for those in the Default Group
- Default Group: If you don't have a grade, make sure you're part of a group (even if it's a group of 1).
- Azure virtual desktop.... If you have problems logging in, create a screenshot and include an explanation of what you tried to do.

Understanding variables vs. functions

- Variables and functions both have names that you type into r without using quotation marks.
- Functions are typically evaluated, and they return a result of some type.
- Functions may return objects like vectors, matrices, data frames etc.

Variables	Functions
<ul style="list-style-type: none">• Can contain any kind of R object<ul style="list-style-type: none">• Single number, vector, data frame, etc• The class() function will tell you what kind of object they hold• They are like nouns in that they don't perform an action, they just represent an object.• Variables aren't followed by parentheses.• We can assign the output of a function to a variable	<ul style="list-style-type: none">• Functions are a particular kind of entity in R.• We type parentheses at the end of a function name to let R know that it's a function and that we want to evaluate it.• Functions are like verbs; they may take an object, and they perform some kind of action, possibly returning a value.

Evaluating a function + saving to a variable

- When we call a function in R, it may return a value or object.
- We can assign the **function output to a variable**.
- For example, in the **expression**:

```
m1 = matrix(1:6, nrow = 2)
```

- First, **matrix()** (a function) is **evaluated**, then **assigns the output to m1** (a variable)
- The material to the right of the assignment operator is always evaluated first.

What's In This Deck?

Slides	Selected Key Take-Home Concepts
<ul style="list-style-type: none">• Data Exploration• Types of Plots• Functions, variables, constants• Formulae and notation• Classes of functions• Intro to distributions	<ul style="list-style-type: none">• Figuring out which parts of a function are variables, and which are constants.• Bases vs. exponents• Exponentials win over powers every time!• Linear, asymptotic, and monotonic.• Summarizing and raw-data plots.

Data Exploration

With examples in built in R!

* R code available on request, absolutely no warranty.

Statistics and Parameters: Frequentist Perspective

We're guests in a Frequentist world

[MS Office Art Suggestion]

- Let's think about data exploration from a Frequentist perspective!
- What is a statistic and what is a parameter?
- What is a population and what is a parameter?



Data Exploration

Numerical	Graphical
<ul style="list-style-type: none">• Compact summary of data• Extremely important, but not as intuitive as a graphical exploration• Summary statistics:<ul style="list-style-type: none">• Center• Spread• 5-number summary	<ul style="list-style-type: none">• Helps you get an intuitive ‘feel’ for what’s in your data.• Graphs/Plots!<ul style="list-style-type: none">• Many types, each shows different aspect of data.• Important distinction: does my plot show all data points, or a summary of aggregated data?

Data Exploration

We're most often interested in two characteristics of our data:

Center

- Mean
- Median
- Mode

Spread or Dispersion

- Range: min and max
- Interquartile range (IQR)
- Variance
- Standard deviation

Data Exploration

- Center and spread are easy to understand numerically.
- Other quantities make more sense graphically:
 - Skew
 - Kurtosis
 - Bi- or multi-modality

Associations

***Association* is a value neutral term.**

- It is useful when you don't want to imply causality, or any specific form of a relationship.

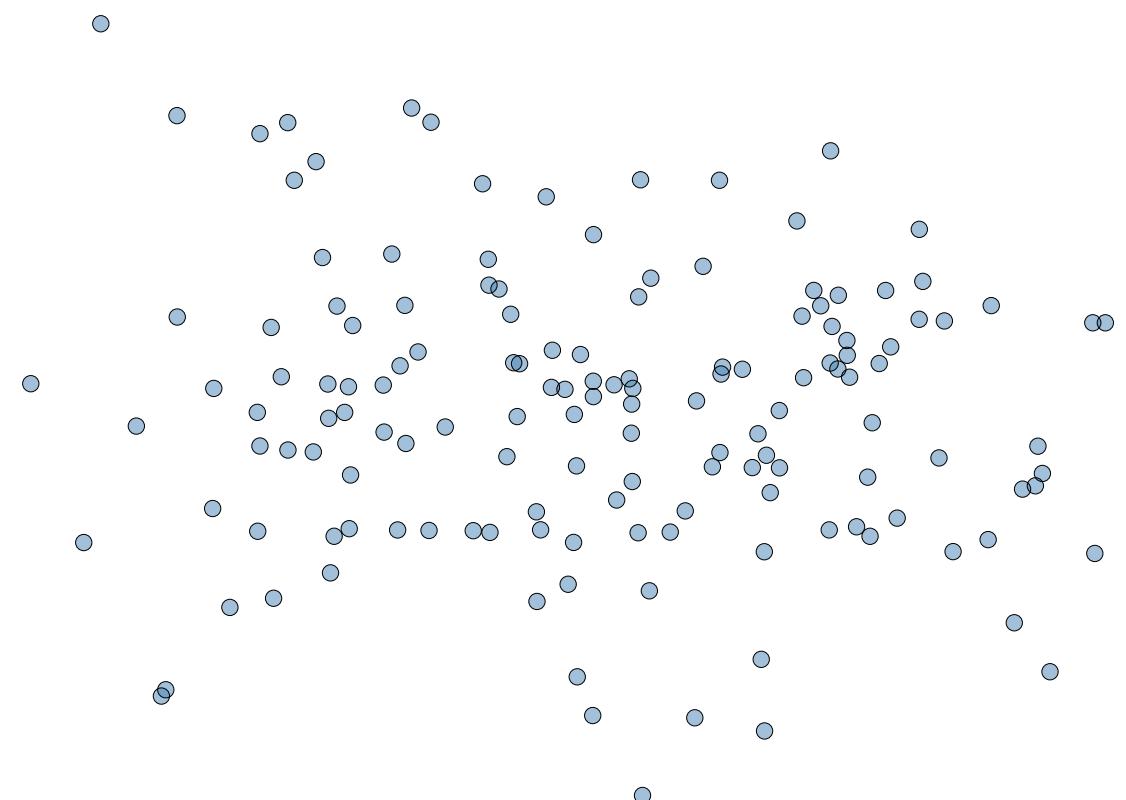
How can we *describe* an association?

- Qualitative and quantitative
- Numerically and graphically

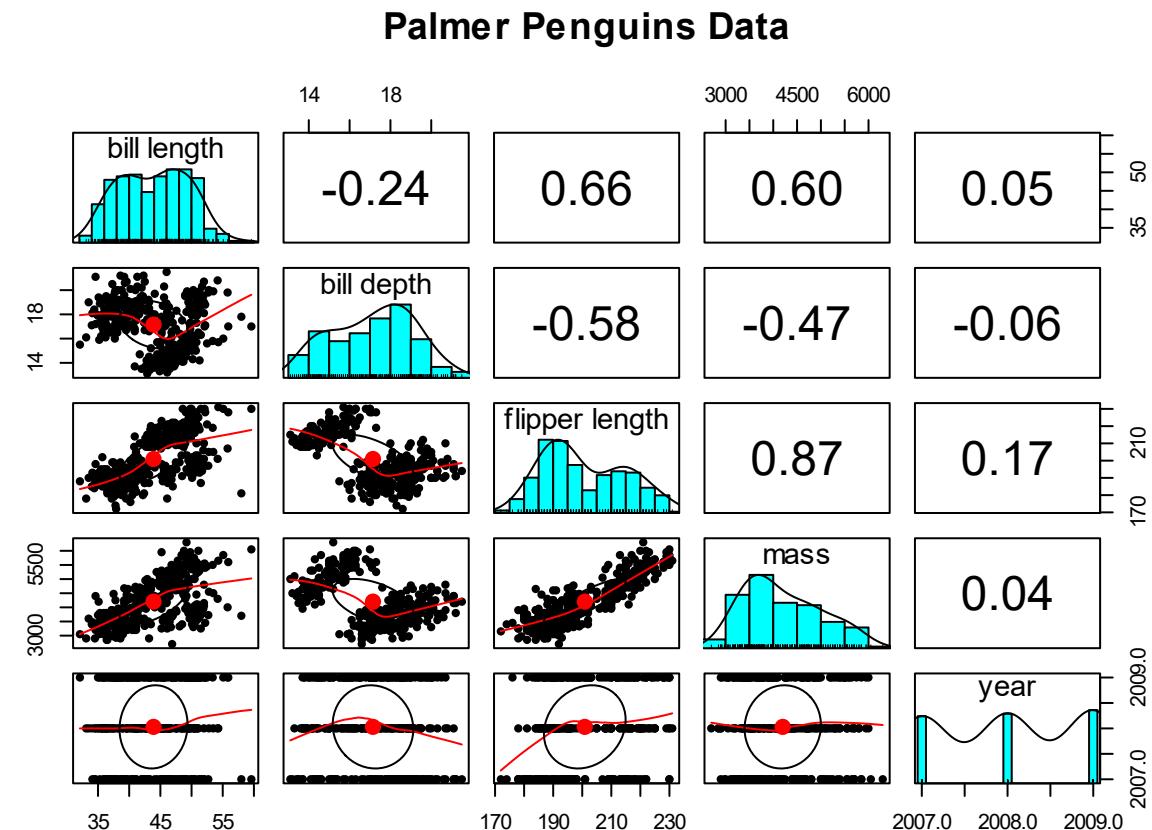


Associations: graphical exploration

Scatterplots are useful



Pairplots are even better!



Correlation Coefficients

Correlations describe the strength of association between two variables

- Correlations measure how close points lie to a curve.
- How well can you predict y from x ?
- Correlations are a kind of descriptive stochastic model.
 - But not a very powerful one, as we'll see.

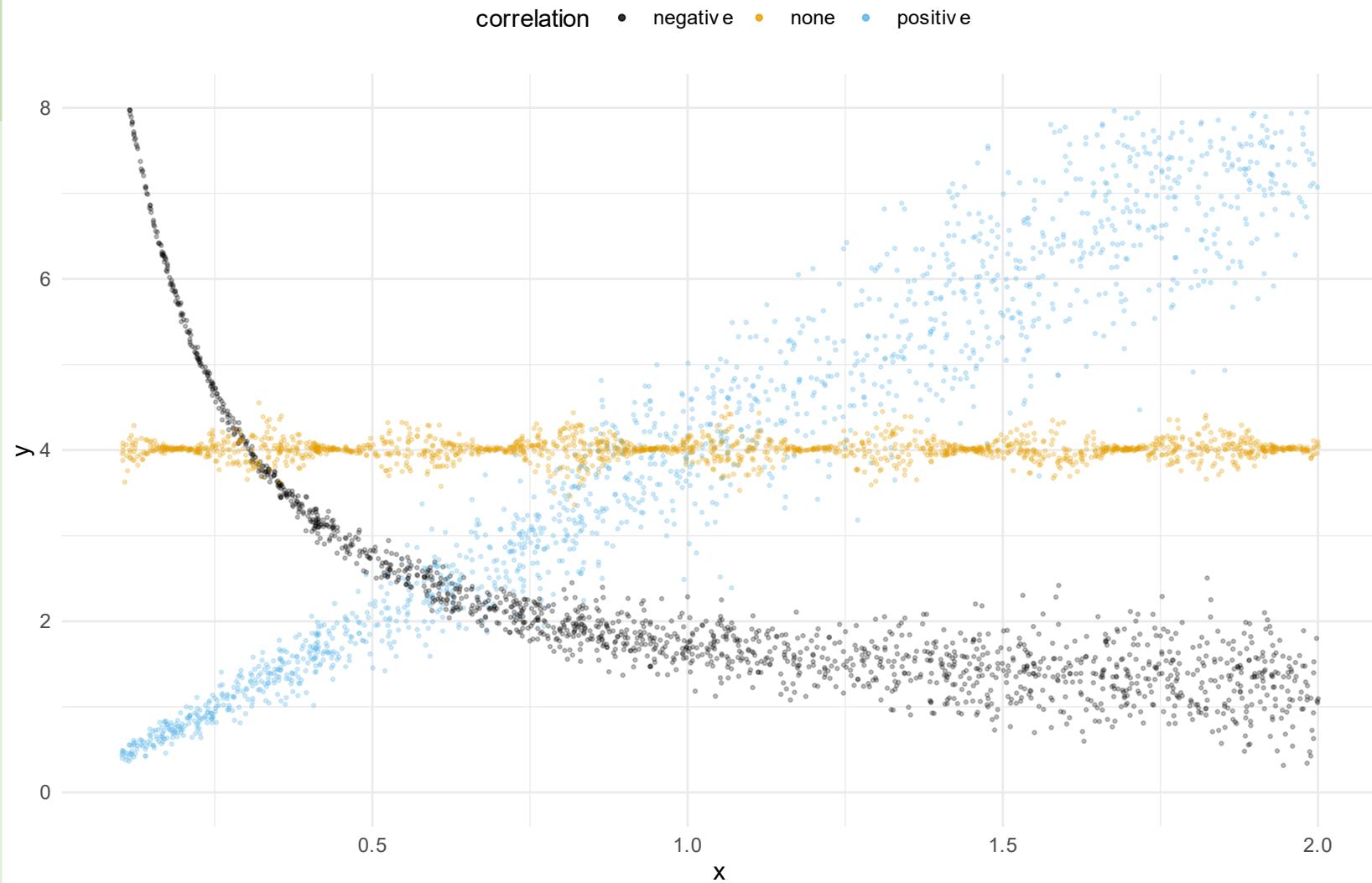
Some limitations and caveats

- Limited to **two** variables.
- Spearman and Pearson correlations are limited to **monotonic** functions.
- Does not tell us anything about the **magnitude** of an association.
- Cannot deal with **multi-collinearity**.
 - But don't worry, we have tools that can.

Correlation

Correlation Measures the Strength of the Association Between two Variables.

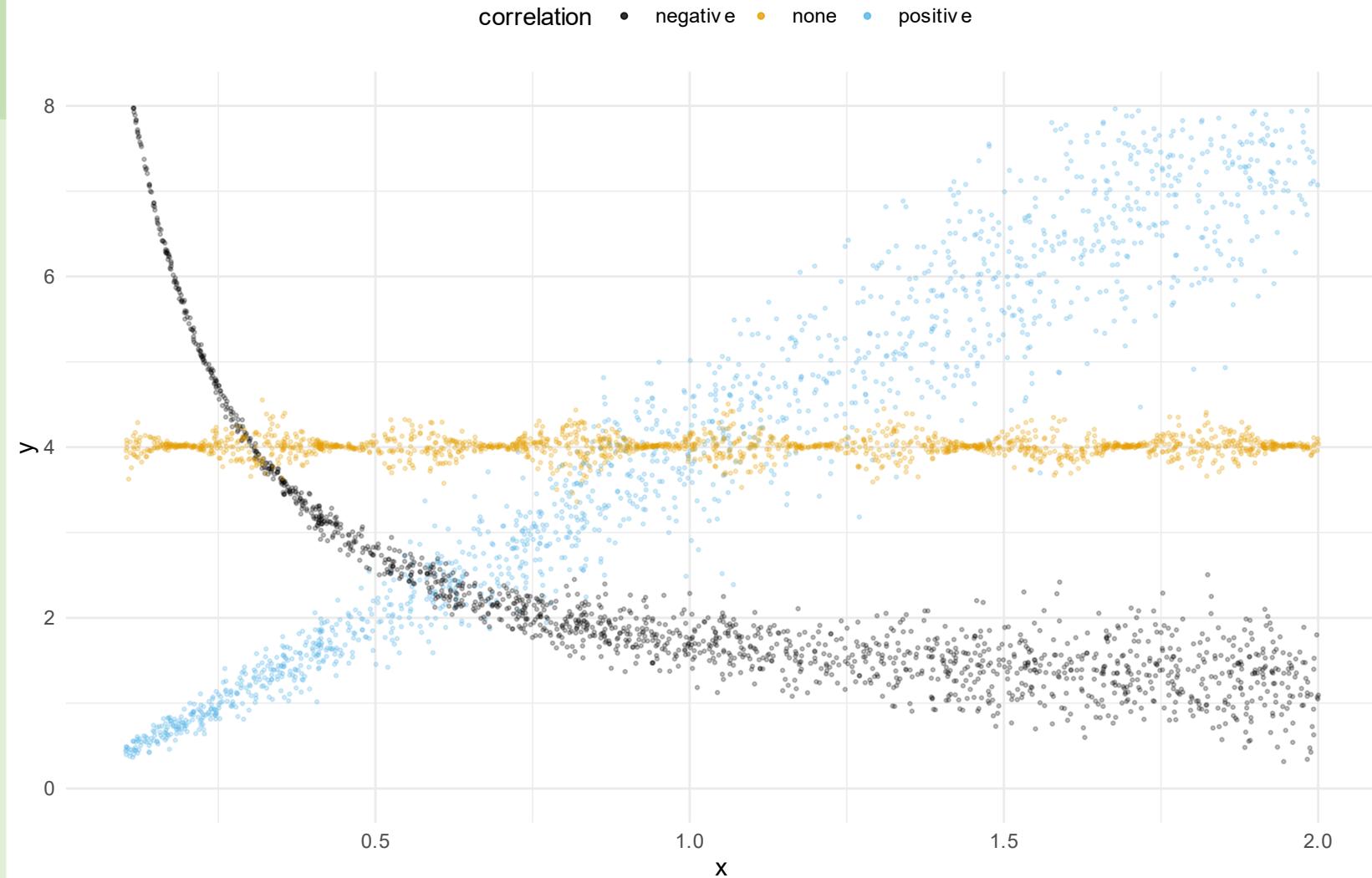
- Correlation ranges from -1 to 1:
- 1 indicates **perfect correlation**
 - Bivariate data lies exactly on a line of positive slope
- -1 indicates **perfect negative correlation**
 - Data lies exactly on a line with negative slope
- 0 Correlation: Points are **totally random** with respect to each variable.



Correlation – Information Perspective

Correlation Measures the Strength of the Association Between two Variables.

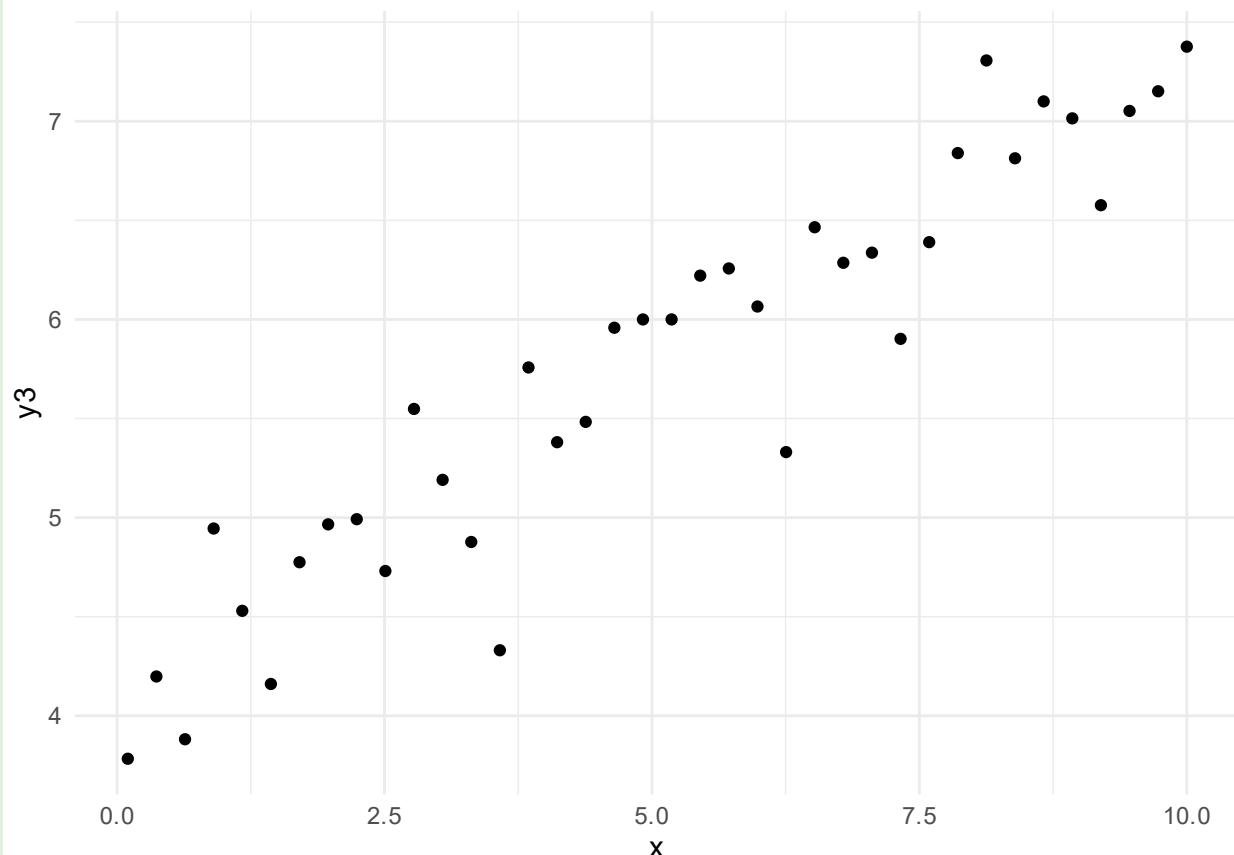
- Correlation ranges from -1 to 1:
- Corr 1: We can predict y from x perfectly. There's no noise, and as x increases, y increases. x tells us all we need to know about y.
- Corr -1: We can perfectly predict y from x, there is a negative relationship with negative slope
- 0 Corr: X tells us nothing about y.



Association: Numerical exploration

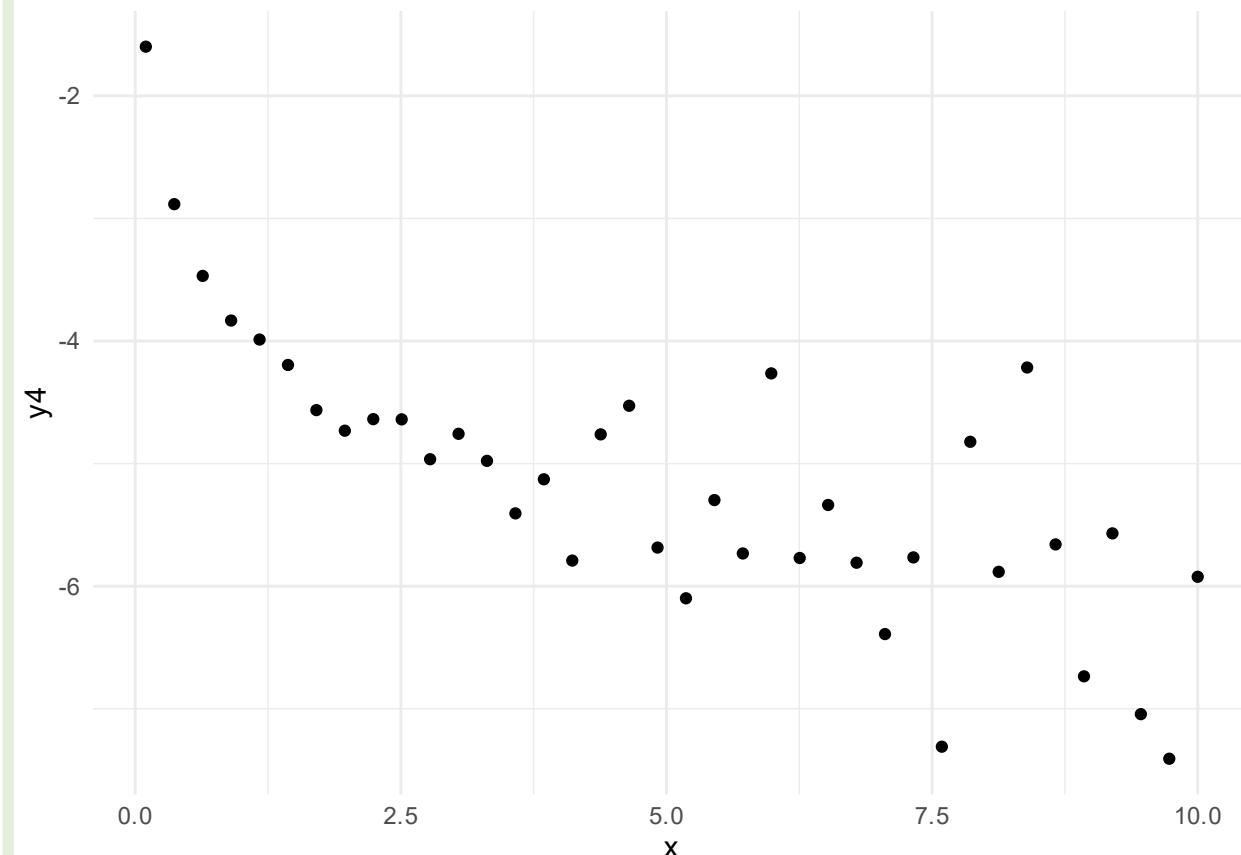
Quantifying linear association Pearson Correlation

Pearson Correlation = 0.879

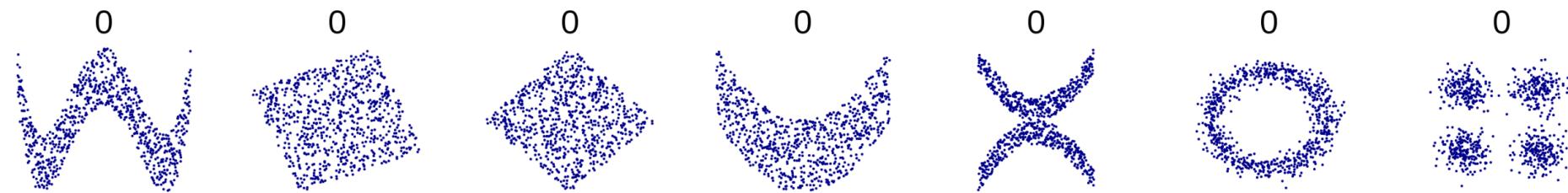


Quantifying monotonic association Spearman Correlation

Spearman Correlation = -0.766



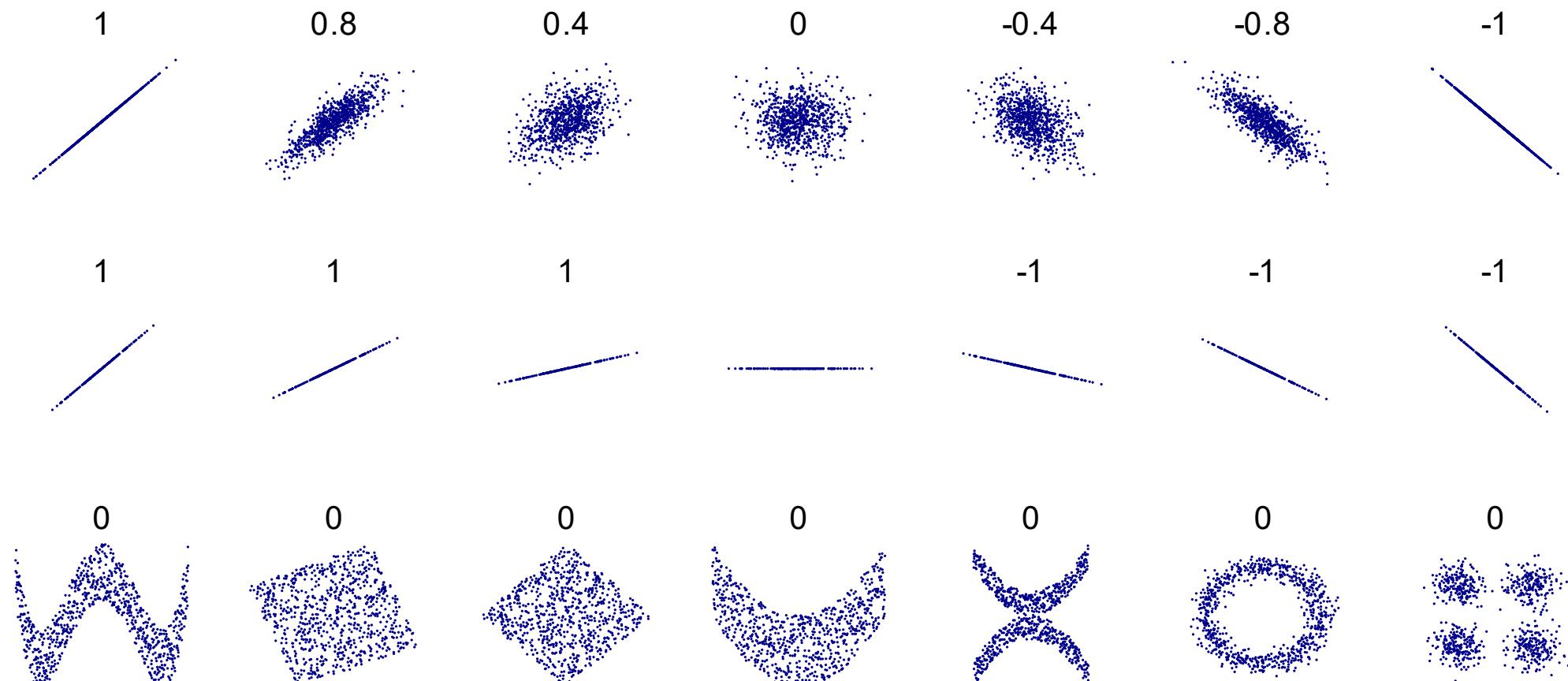
More complicated relationships



- How would you describe these associations?
- Does knowing the value of x tell you anything about y?

- Can you guess what the numbers (all 0) represent?

Correlation Coefficients



Denis Boigelot: public domain image

$s_{xy} = 2$, $r_{xy} = 1$

$s_{xy} = 1.6$, $r_{xy} = 0.8$

$s_{xy} = 0.9$, $r_{xy} = 0.4$

$s_{xy} = 0.1$, $r_{xy} = 0.1$

$s_{xy} = -0.8$, $r_{xy} = -0.4$

$s_{xy} = -1.6$, $r_{xy} = -0.8$

$s_{xy} = -2$, $r_{xy} = -1$

$s_{xy} = 2.1$, $r_{xy} = 1$

$s_{xy} = 1.8$, $r_{xy} = 1$

$s_{xy} = 1.1$, $r_{xy} = 1$

$s_{xy} = 0$, r_{xy} undefined

$s_{xy} = -1.1$, $r_{xy} = -1$

$s_{xy} = -1.8$, $r_{xy} = -1$

$s_{xy} = -2.1$, $r_{xy} = -1$

$s_{xy} = 0$, $r_{xy} = 0$

In-Class R Practice

Announcements: Sep 20th

- Assignment File Formats: pdf or html (occasionally online text)
- Fixed correlation plot for colorblind accessibility (Deck 3)
 - Will work to update others as we go.
- First batch of assignments are graded! Check Moodle and let me or Ana know if there are any issues.
- Software setup assignment: If you didn't get 100%, you may resubmit. This is a special arrangement for this assignment only; it's vitally important to get the software setup correctly.
- In-class R practice feedback.

JOIN THE ECO JUSTICE, EQUITY, DIVERSITY, INCLUSION, AND ACCESSIBILITY (JEDIA) COMMITTEE!

Please join our Leadership Team that meets biweekly throughout the school year, or a Task Force that only meets for a short time period to accomplish a task!

First meeting:

Friday, September 23rd, 11:30-12:30
312A Holdsworth and on Zoom

Email Meg Graham MacLean for more info
mgraham@umass.edu or diverseEco@umass.edu



THEATER
DELTA PRESENTS:

"WHAT'S YOUR PROBLEM?"

An Interactive Theater
performance addressing
conflicts rooted in racism,
sexism, and social identities.

September 21st | Bowker Auditorium
4:00 - 5:30pm & 7:00 - 8:30pm

REGISTER HERE!



Data Dimensionality: How many variables do I have?

Data Dimensionality

What is data dimensionality?

- It's just the number of variables in your data.
- They don't have to correspond to physical and temporal dimensions.
- Axes in 'variable space' or 'parameter space'

Visualizing data

- 1D: boxplots, histograms
- 2D: conditional boxplots, scatterplots
- 3D: coplots, 3D plots, 'slices'
- 4D is difficult or impossible
 - Multiple 3D 'panels'
- 5D and higher is generally impossible with single plots.

Scatterplots and Pairplots

Scatterplots

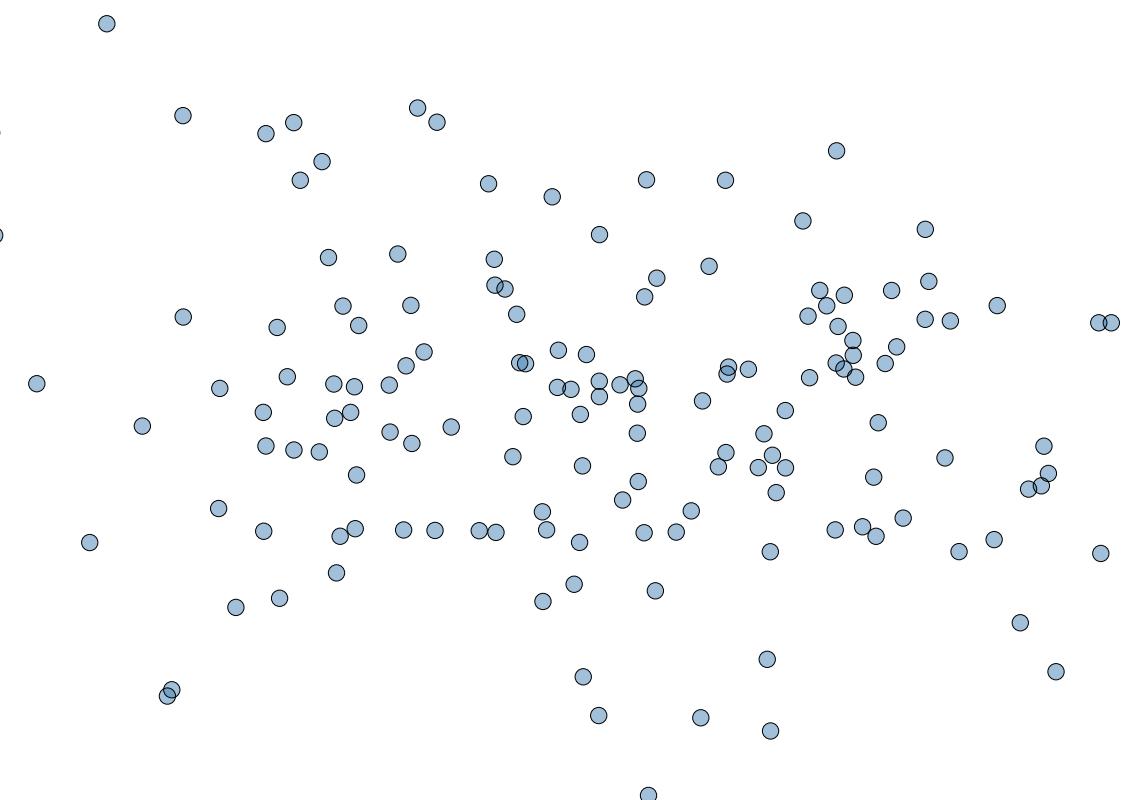
- Good for 2-dimensional data
- Requires 2 continuous variables
- 3D scatterplots are possible, but they can be hard to interpret if they're not interactive.

Pairplots

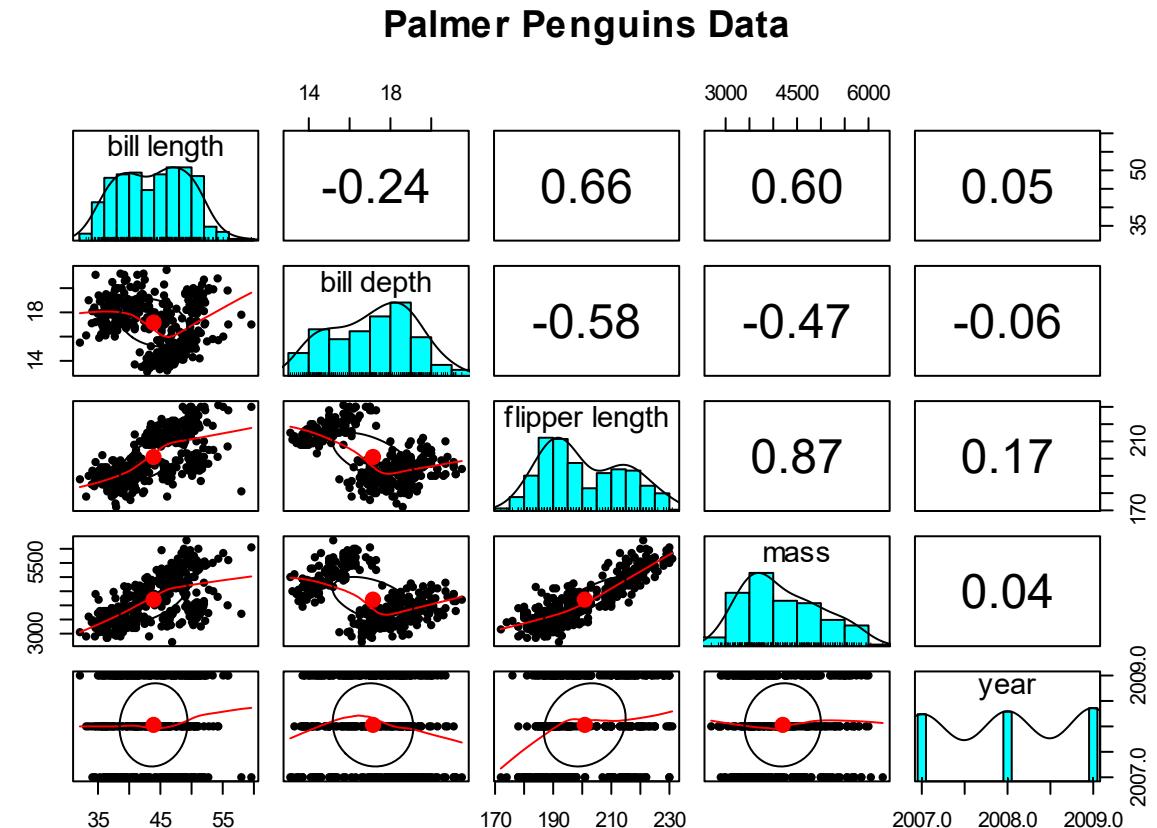
- Matrix of 2D scatterplots.
- Useful for multi-dimensional data.

Scatterplots and Pairplots

Scatterplots are useful



Pairplots are even better!



Visualizing 3D Data: Coplots

Visualize 3-Dimensional data with 2D slices

- Individual data points are plotted on x-y plane
- The z-axis is divided into bins
- Straightforward for categories
- Binning algorithm needed for continuous
- Each z-bin is flattened and plotted as 2D



Visualizing 3D Data: Coplots

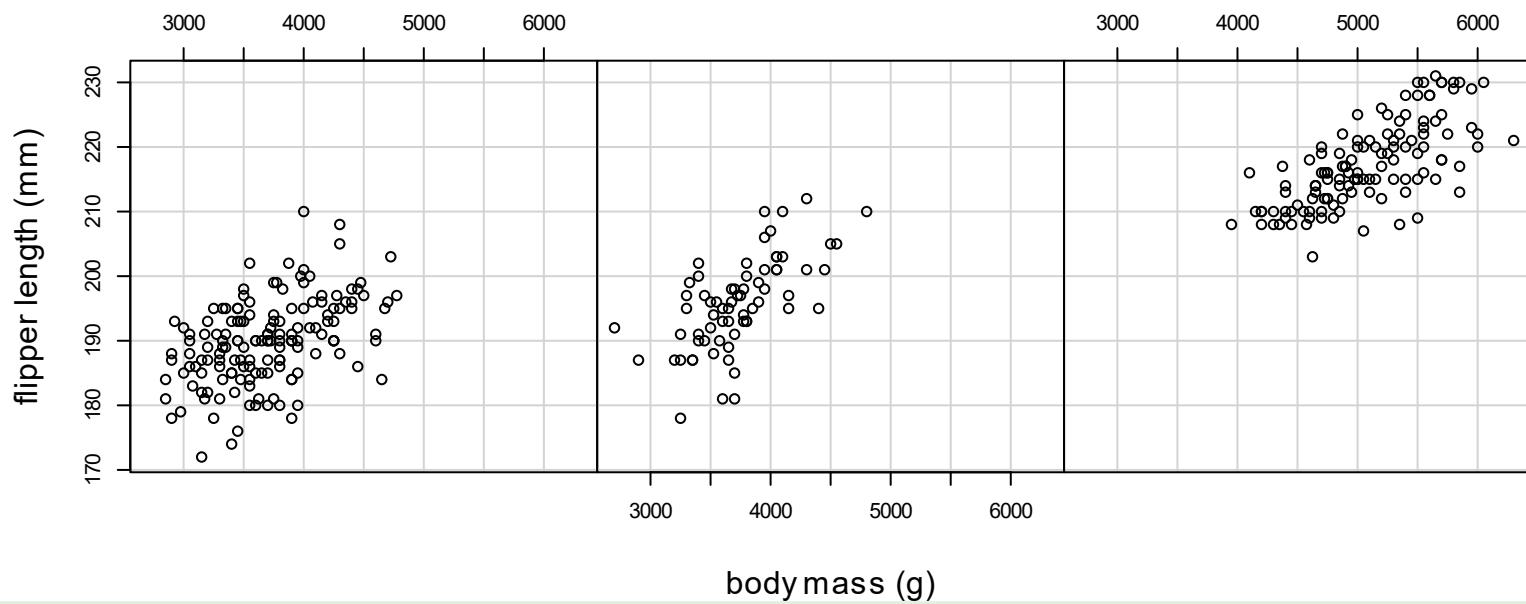
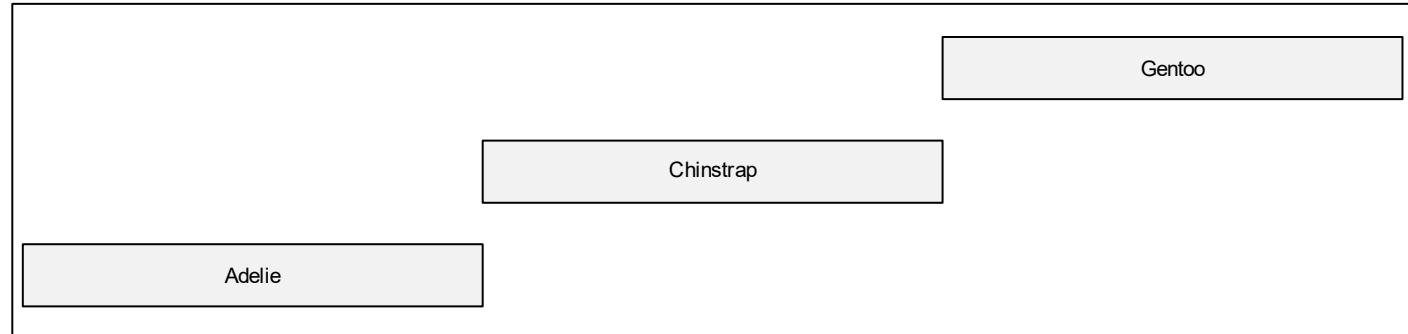
Coplot with a categorical variable

Each slice is a penguin species:

- Adelie
- Chinstrap
- Gentoo

What can you see?

Given : species



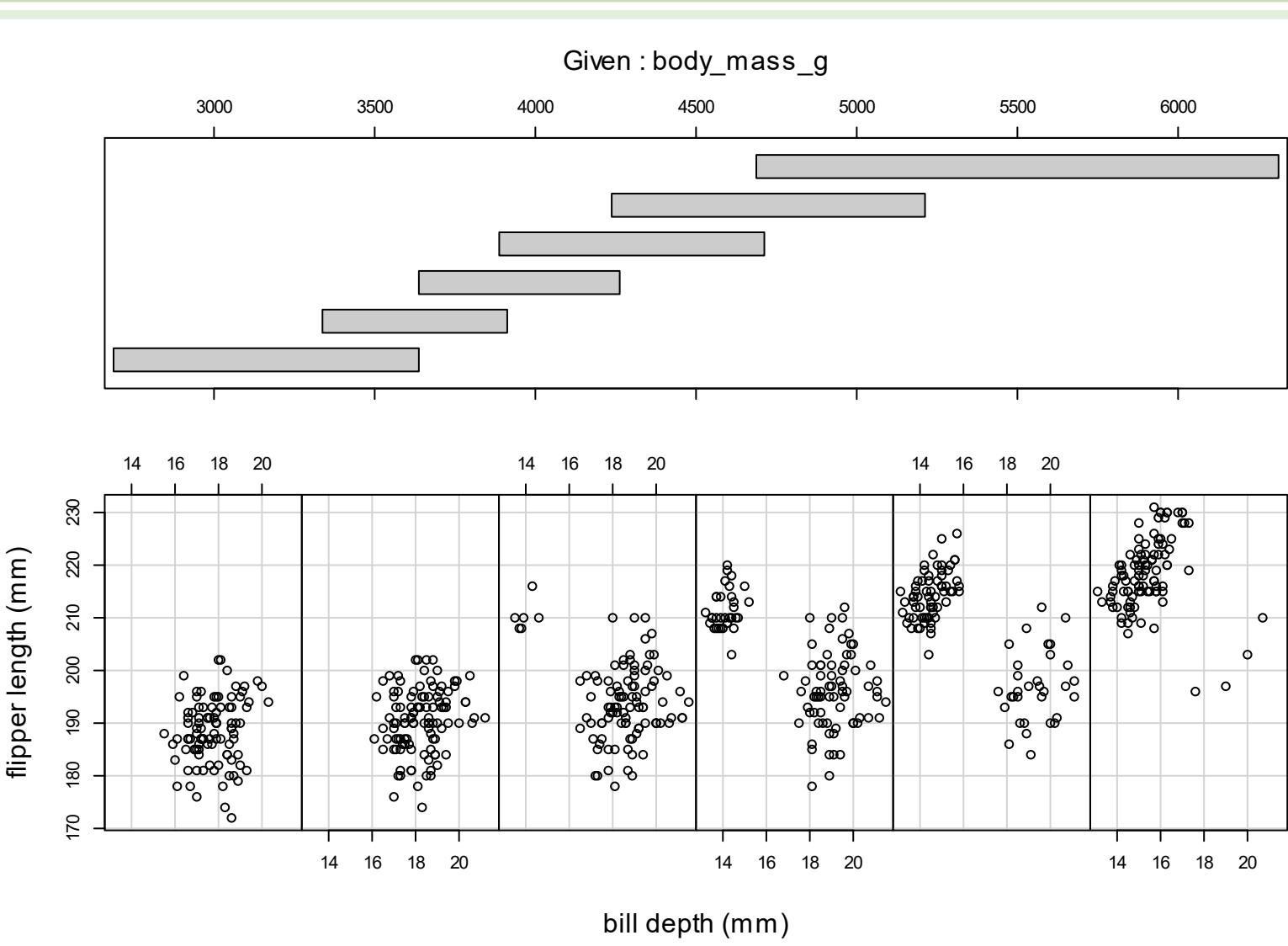
Visualizing 3D Data: Coplots

Coplot with a numeric variable

Body mass broken into 6 'bins'.

What insight does this plot show?

Can you explain the two clusters at greater body mass?



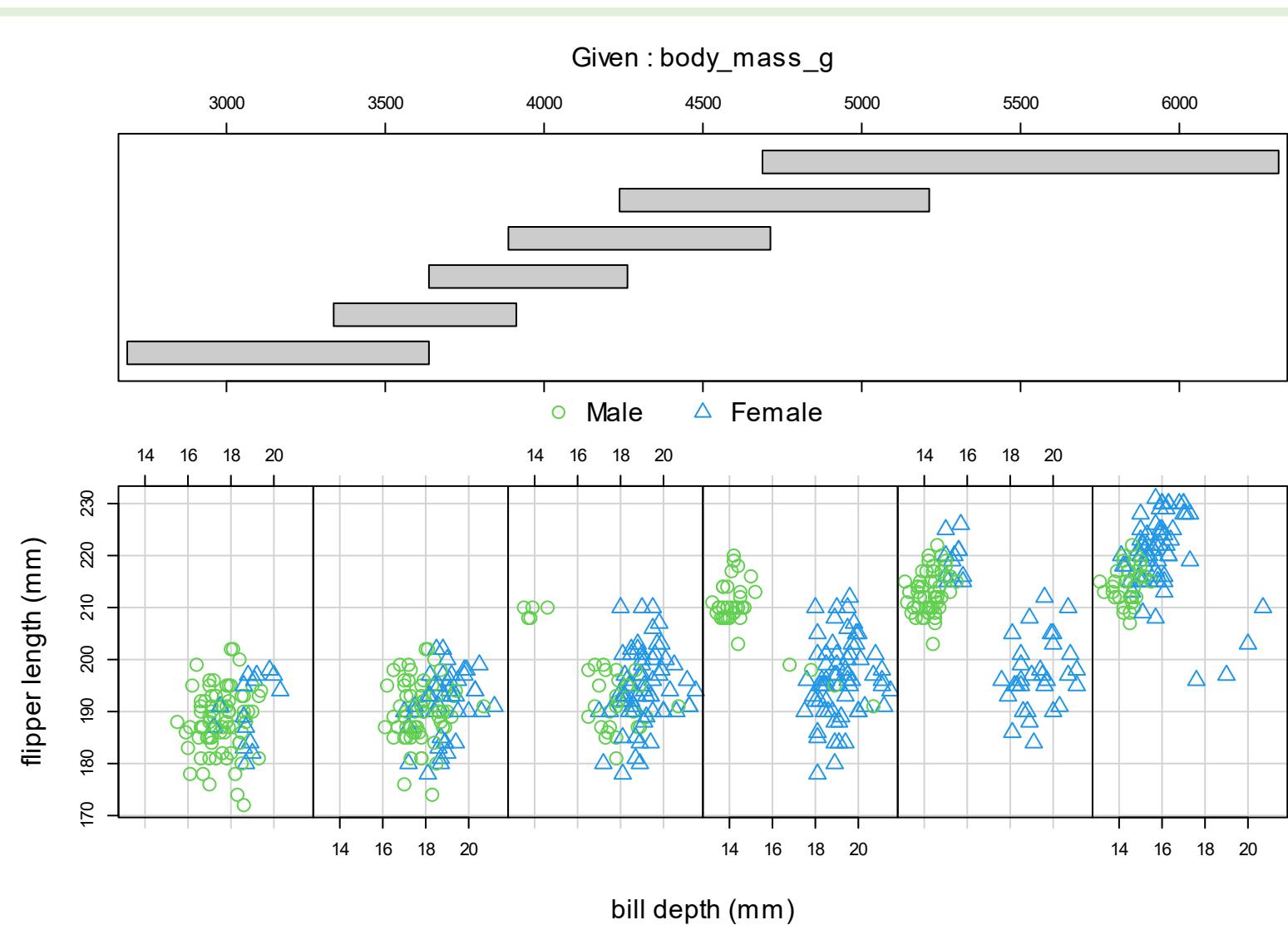
4D Slice Plots: Point Color and Shape

Coplot with a numeric conditioning variable and 4th dimension as point shape.

- 6 body mass bins
- Sex as plotting character

Do the groups make more sense now?

- What factor(s) is/are still missing?
- How could you put this into an English sentence?



3D example: Modeling Mountain Pine Beetle epidemics

3-dimensional data

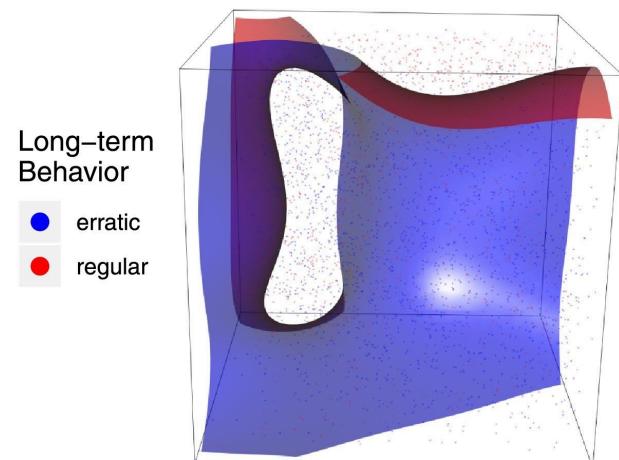
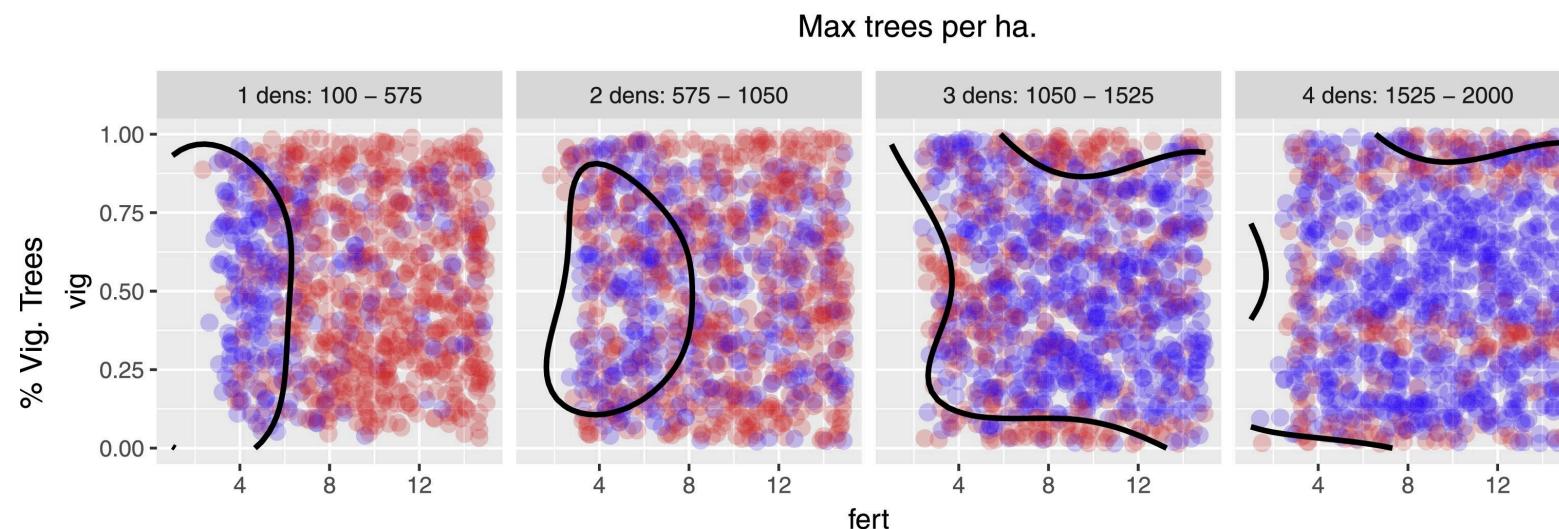
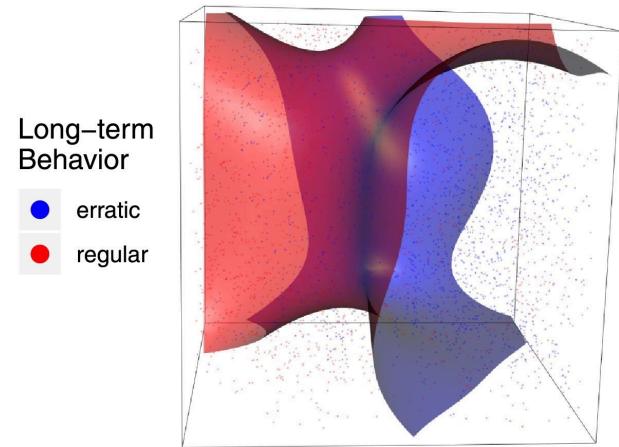
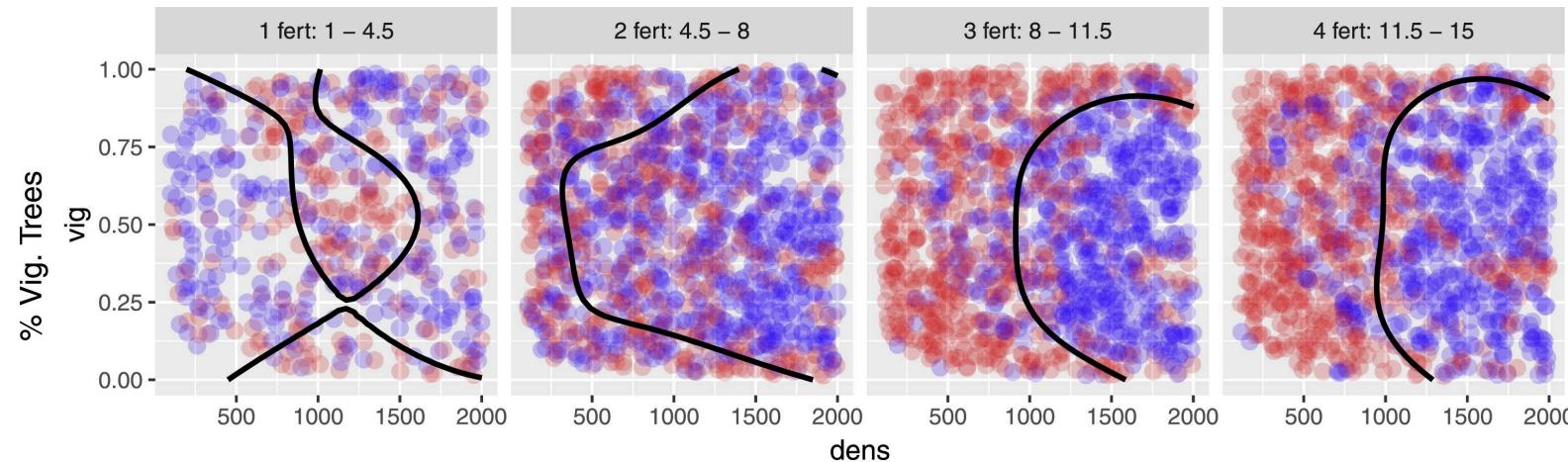
Three model parameters:

1. Beetle fertility
2. Tree vigor
3. Tree density

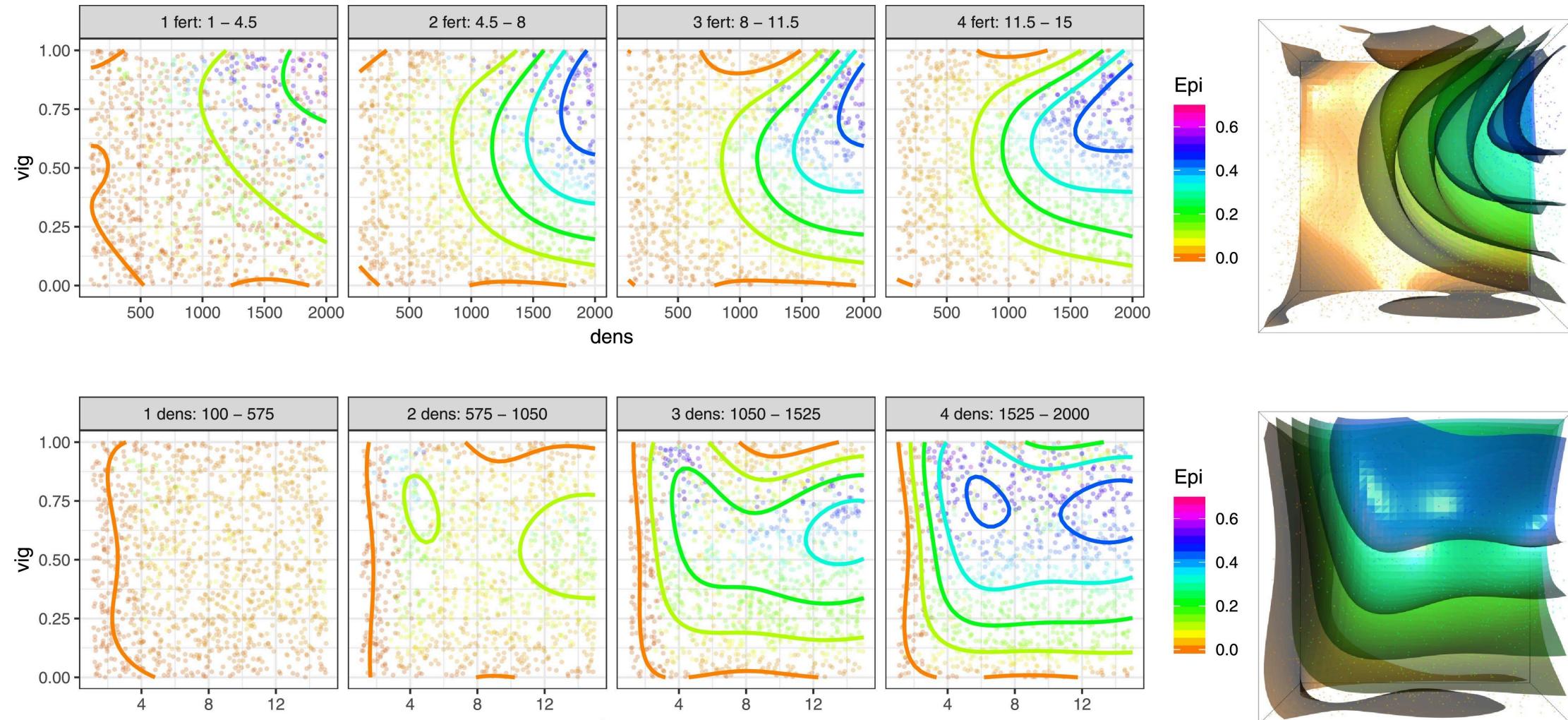
Two response types:

1. Long term epidemic behavior - categorical: erratic or regular epidemics
2. Epidemic proportion - continuous: long-term average percent of epidemic area

2D Slice Plots: Slices + Epidemic Behavior (discrete) Response as Color

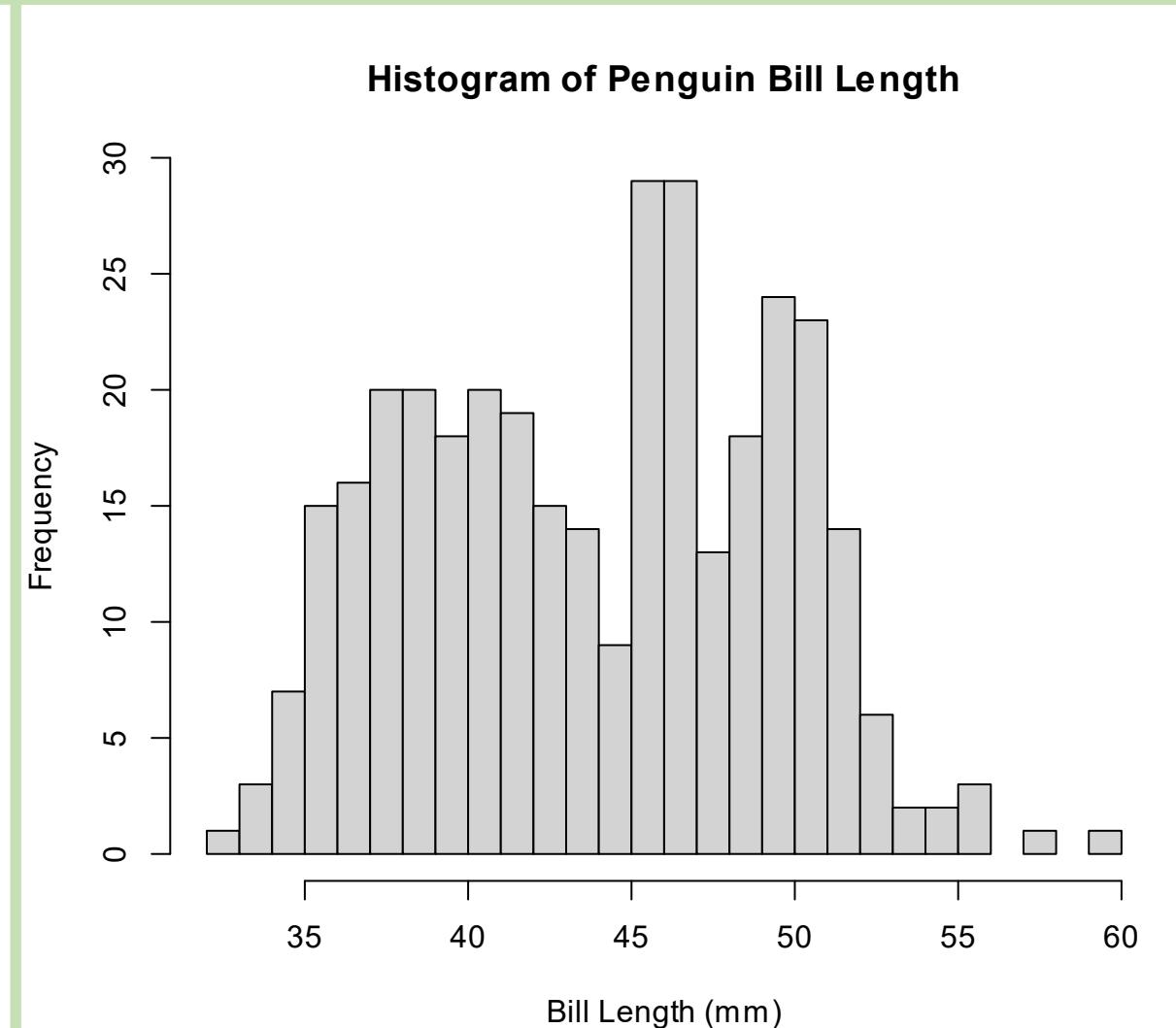


2D Slice Plots: Slices + Epidemic proportion (Continuous Response) as Color



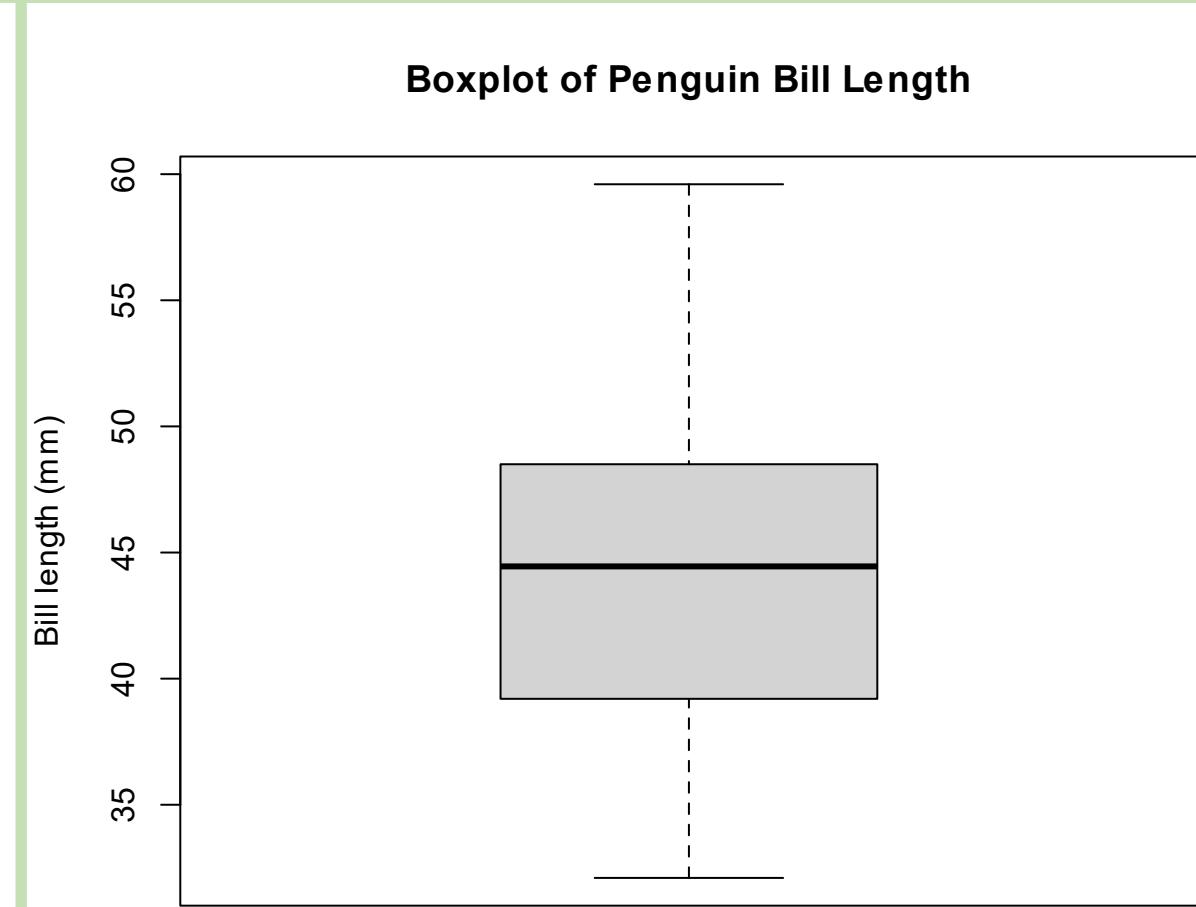
Univariate Data: Histograms

- Show the ‘shape of the distribution’ of the data.
- Sometimes called frequency diagrams.
- Data are aggregated into bins



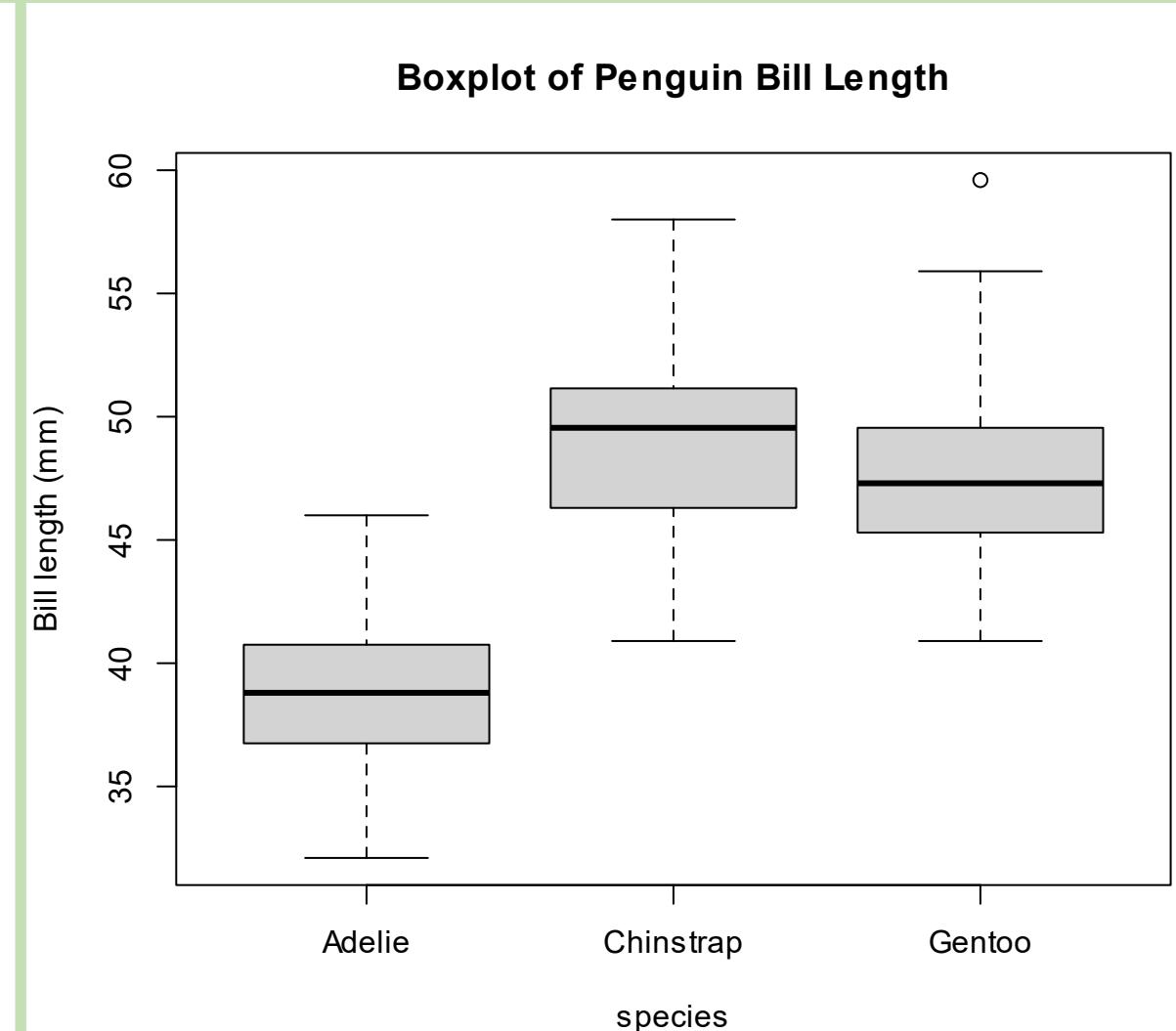
Univariate Data: Boxplots

- Summarizes data: like a 5-number summary.
 - IQR, Median
 - Whiskers: it's complicated
 - <https://www.r-bloggers.com/2012/06/whisker-of-boxplot/>
- Univariate: but can be made bivariate, or even 3D with grouping factors.
- Gives you an idea about spread and skew.
 - How is it different from a histogram?



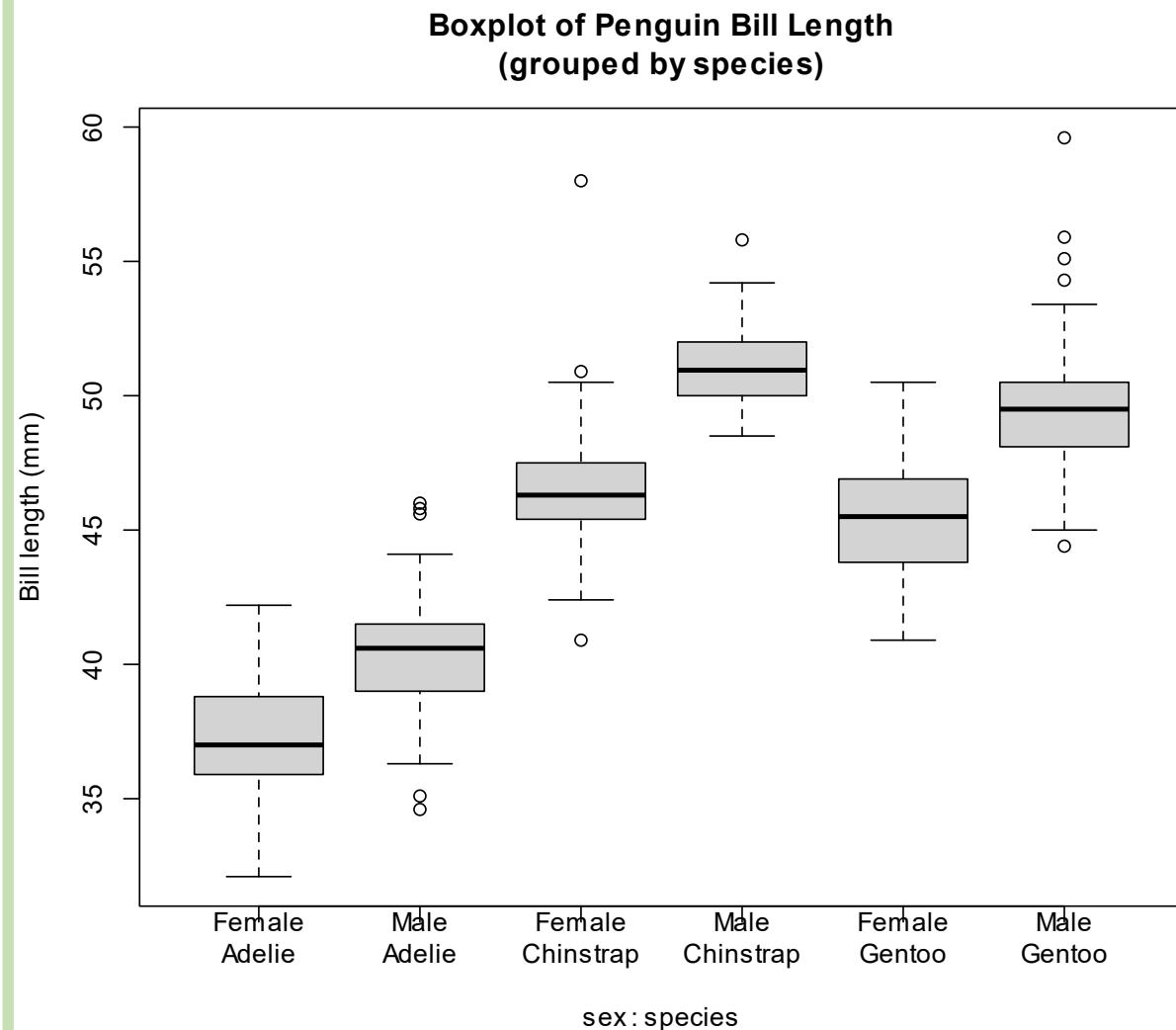
Conditional Boxplot: Univariate with Grouping Factor

- Univariate Data, aggregated by a grouping factor.
 - Kind of like a coplot
- Summarizes data just like a regular boxplot



Doubly Conditional Boxplot

- Two grouping factors
- 3D data



Graphical Exploration Recap

- Associations
- Tools to describe associations
 - Spearman and Pearson correlation
 - Graphical exploration
- Data dimensionality and plotting
 - Bivariate and univariate plots
- Plots that show all data: coplots, scatterplots, 3D scatterplots, others?
- Plots that show aggregated data: histograms, boxplots, others?
- Grouping factors (conditioning variables)

In-Class Data Exploration

- NOTE: Moodle group self-select will be slightly different going forward!
 - Setting was wrong on last Thursday's assignment
 - Use the group self-select for every assignment going forward.
 - Group names will have to be slightly different each time (it's a pain, but it's the best I can do to keep track of groups...)
 - Feel free to shuffle groups as you wish, that's why I have a different group/grouping set up each time.

Announcements

- Re-load the main course page often!
- Remember office hours: great time to ask questions.
 - Tuesday/Thursday 1-2
 - By appointment (email me!)
- Echo360 Lecture recordings.
- Several assignments are due on the 25th:
 - Check the ‘Upcoming events’ feature on Moodle to see the timing of due dates.

COLLEGE DAY COOKOUT

Time: 11 a.m. – 1 p.m.

Date: Friday, September 23

Location: Integrated Science Building
(ISB) Courtyard!

EAT GREAT FOOD!

MEET NEW FRIENDS!

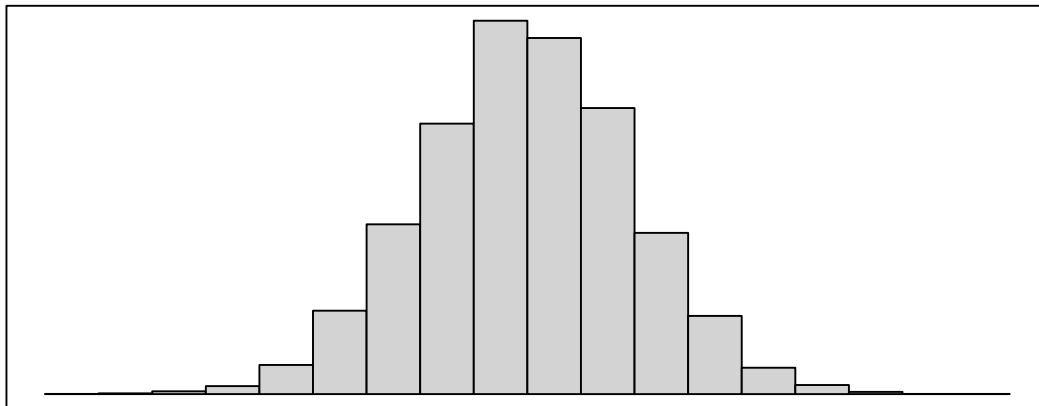


UMassAmherst

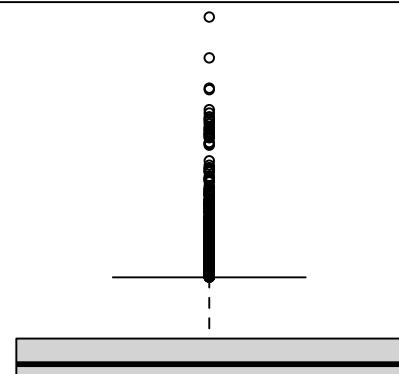
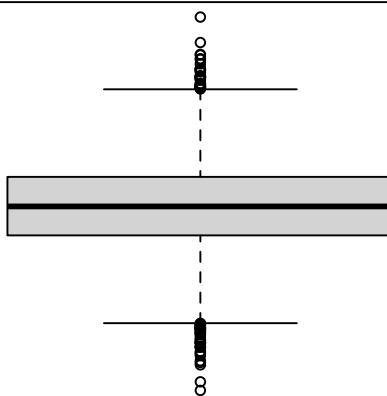
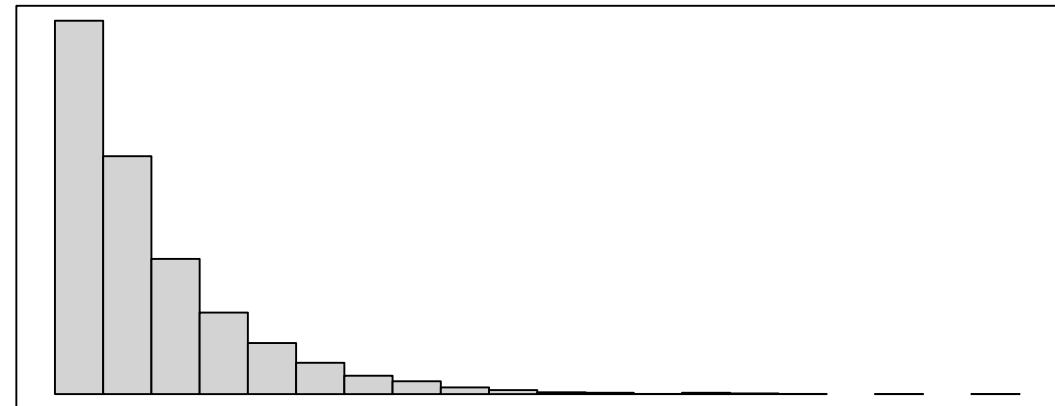
College of
Natural Sciences

Symmetrical and Skewed Data: Histogram and Boxplot

symmetrical



skewed

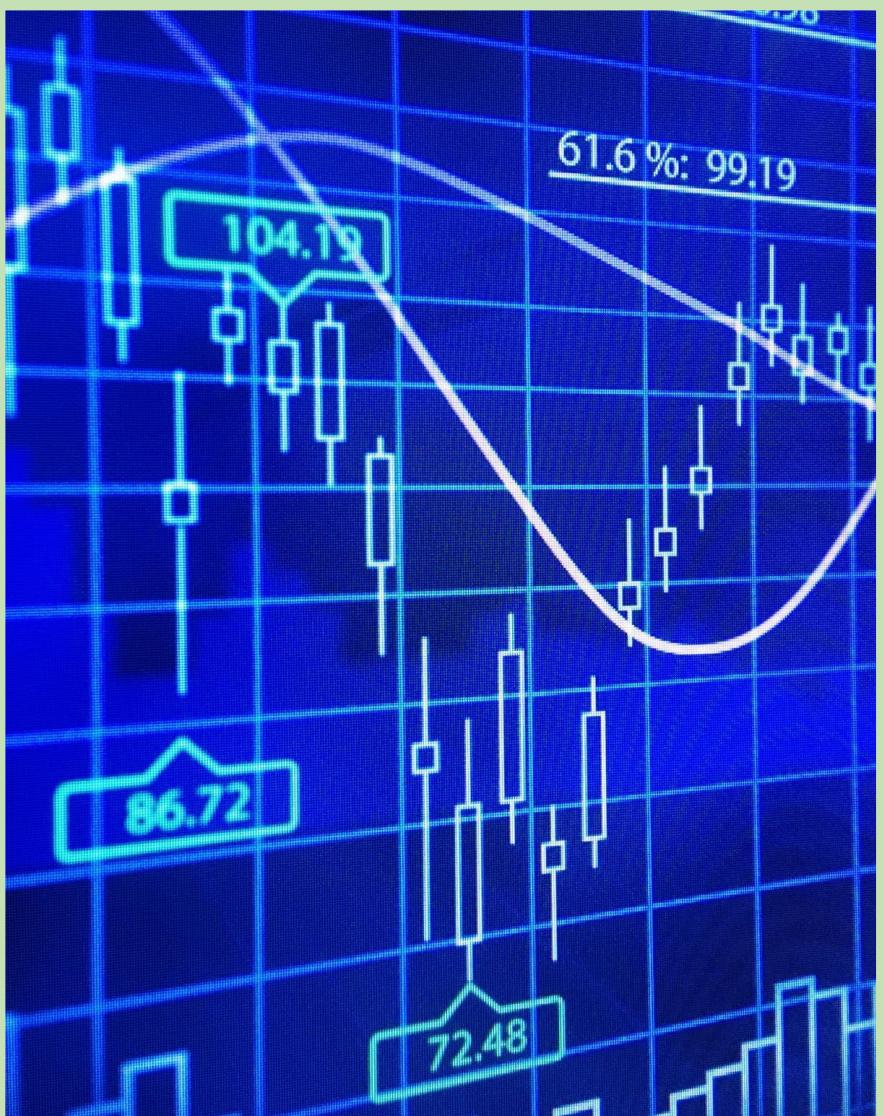


Function basics: important function terms

- **monotonic, asymptotic, and divergent**
- **variables and constants**
- **powers and exponents**
- **local linearity**
- **domains: bounded and unbounded**
- **sums and integrals**
- continuity, slope, and step functions
- saturating, diminishing returns
- inverses



variables and constants



- Constants may also be referred to as parameters

How many variables are there in this equation?

$$P(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- A good strategy to simplify the form of a function is to set all constants to zero or one. That way you can eliminate them from the formula, leaving only the variables.
- Hint: How many times does x occur?

Variables and constants, i.e. parameters

How many variables are there in this equation?

- Only 1: x
- Use the strategy above to eliminate the constants:

$$P(x) = e^{-x^2}$$

That is considerably simpler to understand than the original monstrosity!

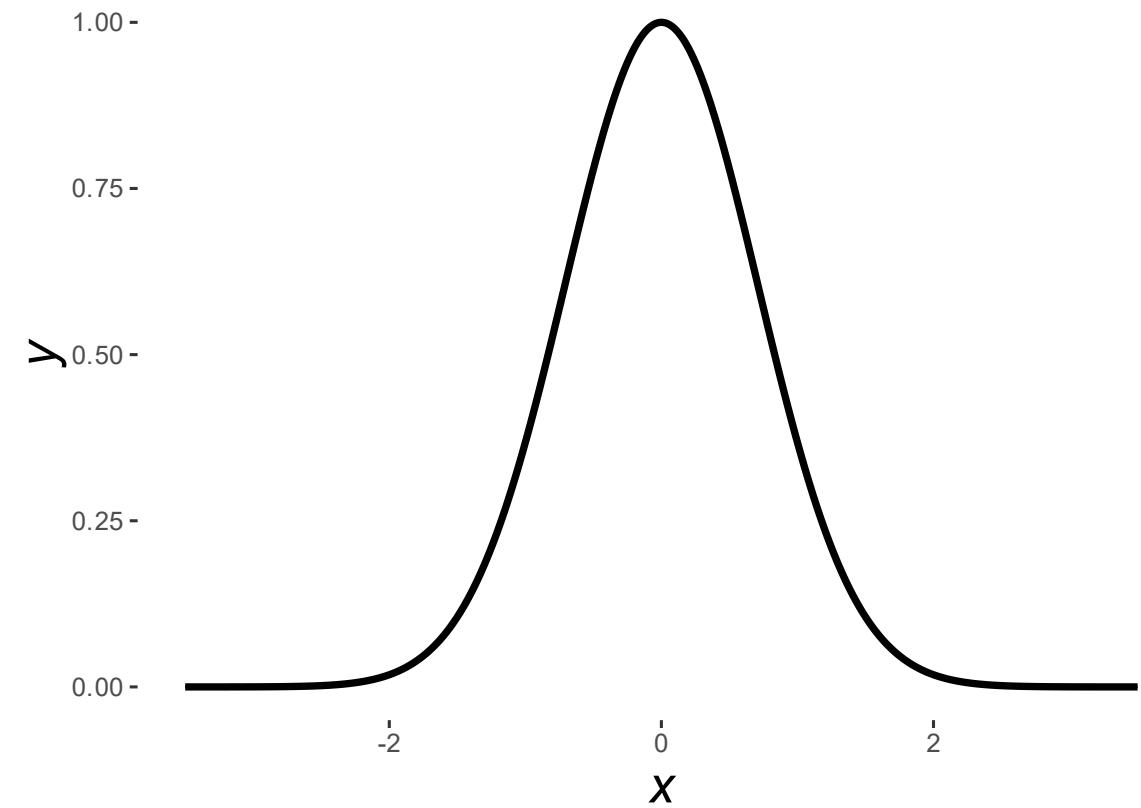
Pi, etc. sind parameter/constants

$$P(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Variables and constants, i.e. parameters

We can plot this: it looks like the Normal curve!

$$P(x) = e^{-x^2}$$



Class of functions

Common functions we use as deterministic models:

- linear
- polynomial, rational, and power
- exponential (and logarithmic)
- periodic
- combination functions

Deterministic = average behaviour of data,

Dual paradigm = deterministic + stochastic model

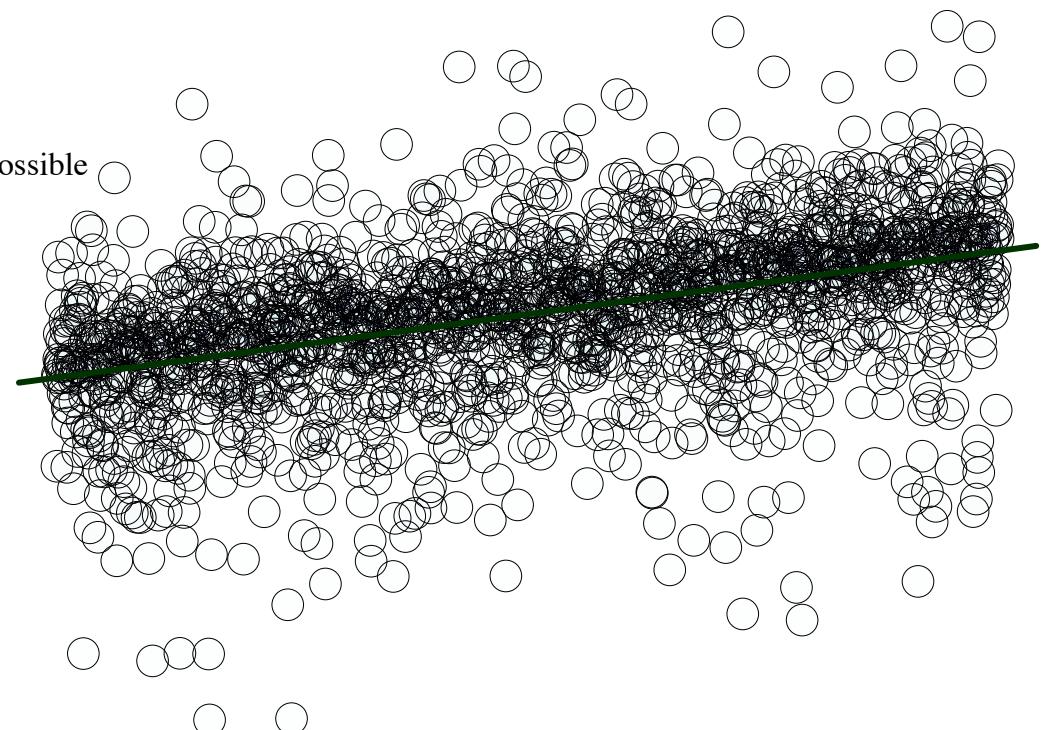
deterministic: function line

stochastic: reality distribution around it

Linear functions

Linear functions have variables raised to a power of 1.

A linear model is always a great place to start.



Linear models: interpretation

A linear model describes a *constant rate of change*

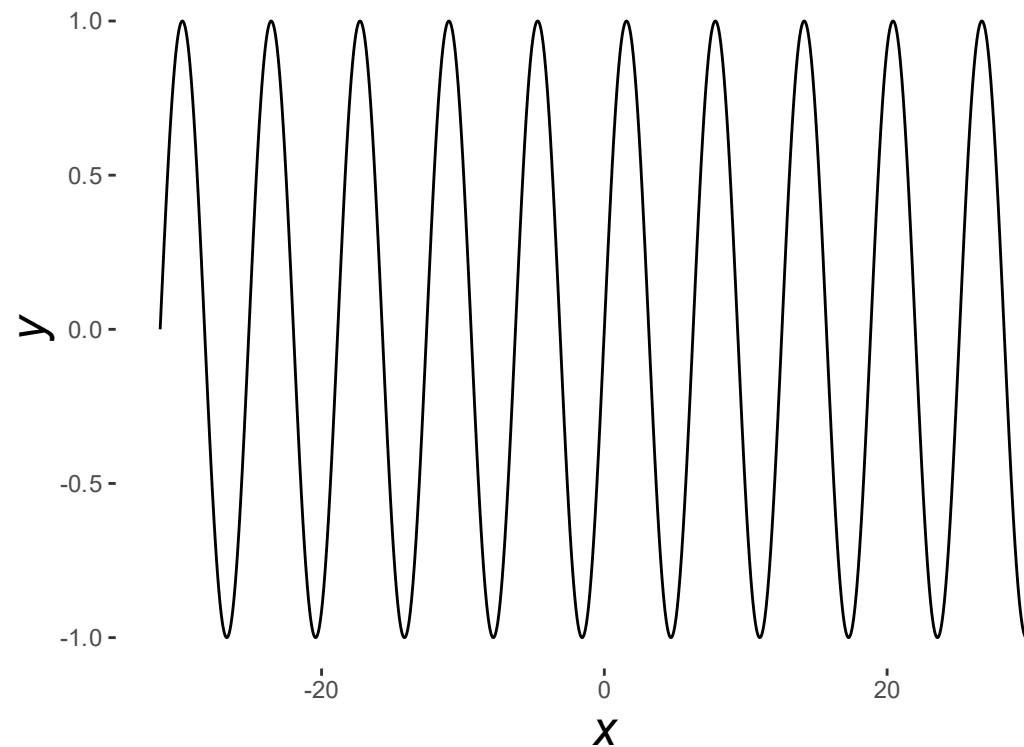
- Reed canary grass plant biomass increases by 0.71 grams for each additional gram of added soil nitrogen per cubic meter.
 - The rate of increase is constant everywhere:
 - If soil with 1g nitrogen results in biomass of 1 g.
 - Soil with 2g nitrogen: expected biomass: 1.71 g.
 - If soil with 3000g nitrogen results in biomass of 100 g.
 - Soil with 3001g nitrogen: expected biomass: 100.71 g.

Constant rate of change: slope always gonna be the same, same steepness

Is the constant rate of change reasonable?

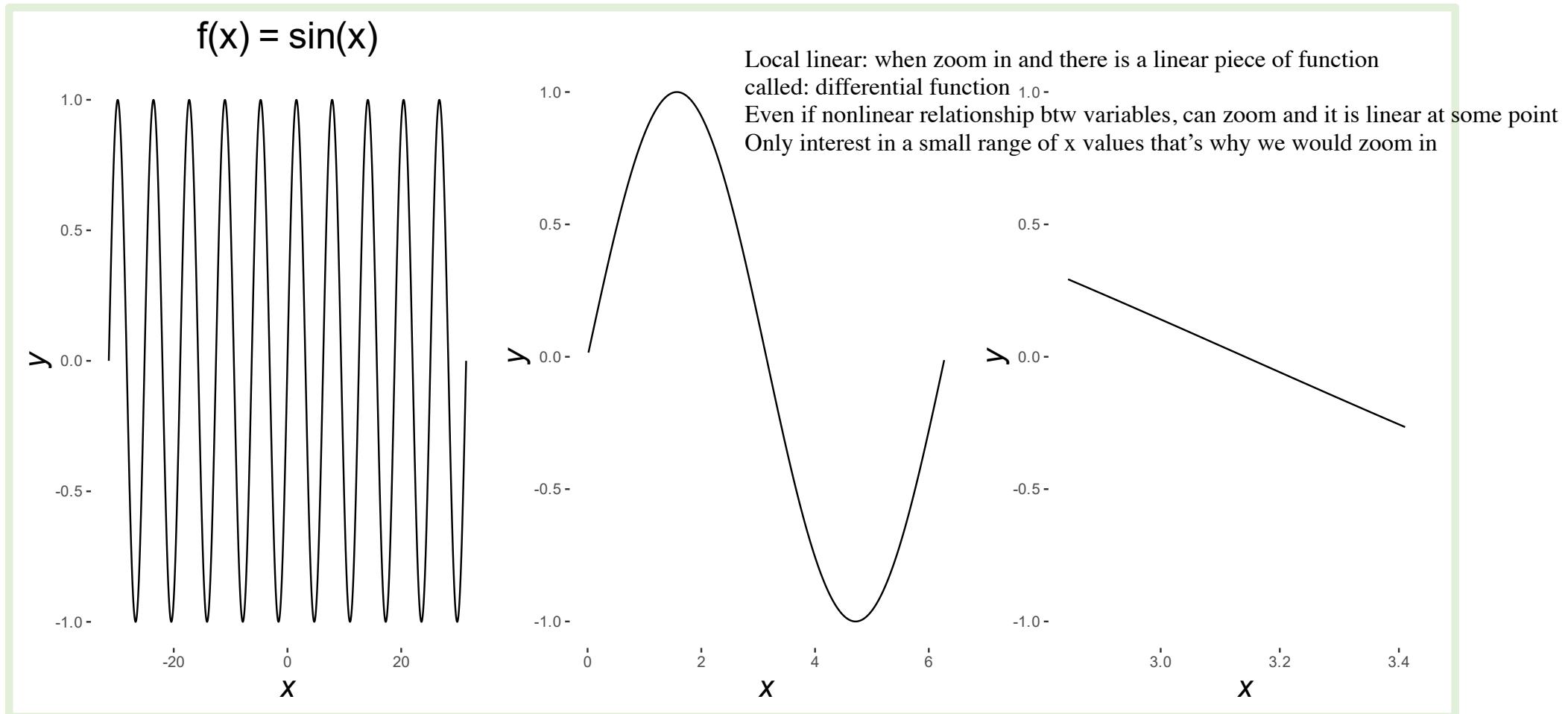
Local linearity

Could you model this curve with a linear function?



Local linearity

Could you model this curve with a linear function?



Local linearity

We are often justified in using a linear model, even when we know a relationship isn't linear:

- If we are only interested in a small subset of the range of predictor values.
- All (most) continuous functions look very linear if you zoom in.
- Linear functions are much simpler than the rest of the functions we'll consider.

Atlas of Function Classes

A selection

Notation for Polynomial Functions

What are bases and exponents?

Let a be a real number, and b be an integer:

a is the base

a can be any number, b is integer

b is the exponent

Exponentiation in this world means:

“Multiply a by itself b times.”

Notation convention:

a^b

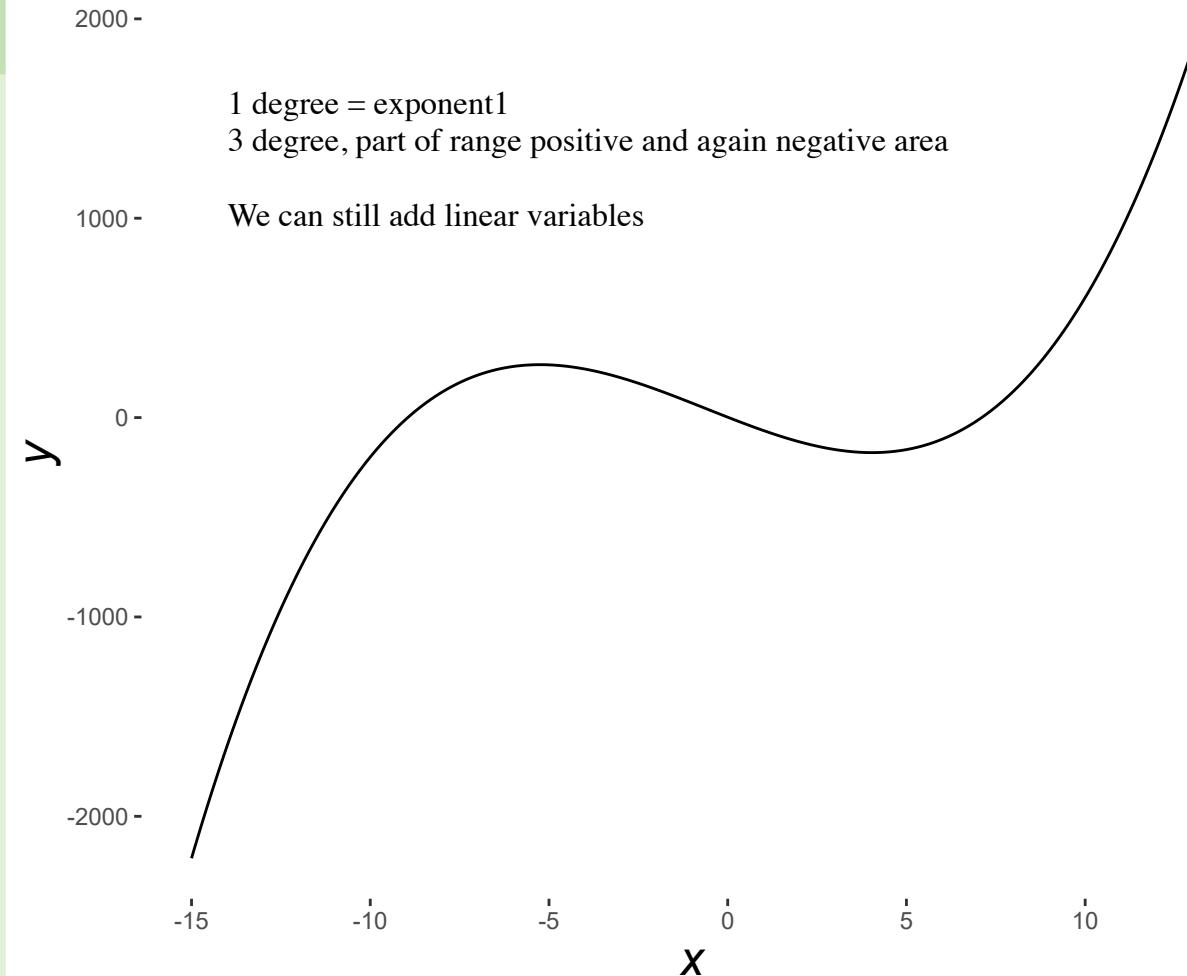
Polynomial functions

Polynomial functions have non-negative integer powers:

$$f(x) = x \quad f(x) = x^3 - 2 \times x^2$$

Linear functions are a subset of polynomial functions.

$$f(x) = 1.1x^3 + 2x^2 - 70x + 2$$



Polynomial models

Polynomial terms are sometimes added to models to improve the **model fit**.

- Polynomial models are typically *phenomenological*.
- There's usually not a clear biological or ecological interpretation.
- You can think of them as *tuning* parameters to increase model fit, or to help with normality of the residuals.

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$$

- Notice this polynomial model is *linear* in the *parameters*!

What does ‘linear in the parameters’ mean?

Polynomial as deterministic part of the model can be used to find line of best fit in function
Don't use them to understand system
Have data that look not linear can fit them into polynomial function

Polynomial models

Polynomial terms are sometimes added to models to improve the **model fit**.

- Polynomial models are typically *phenomenological*.
- There's usually not a clear biological or ecological interpretation.
- You can think of them as *tuning* parameters to increase model fit, or to help with normality of the residuals.

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$$

- Notice this polynomial model is *linear* in the *parameters*!

What does ‘linear in the parameters’ mean?

- The parameters (the betas) are not bases or exponents.

Rational functions

Rational functions can be expressed as a ratio of *polynomial* functions.

Rf variable is raised towards non integer power

X raised to negative 1 power = 1/x

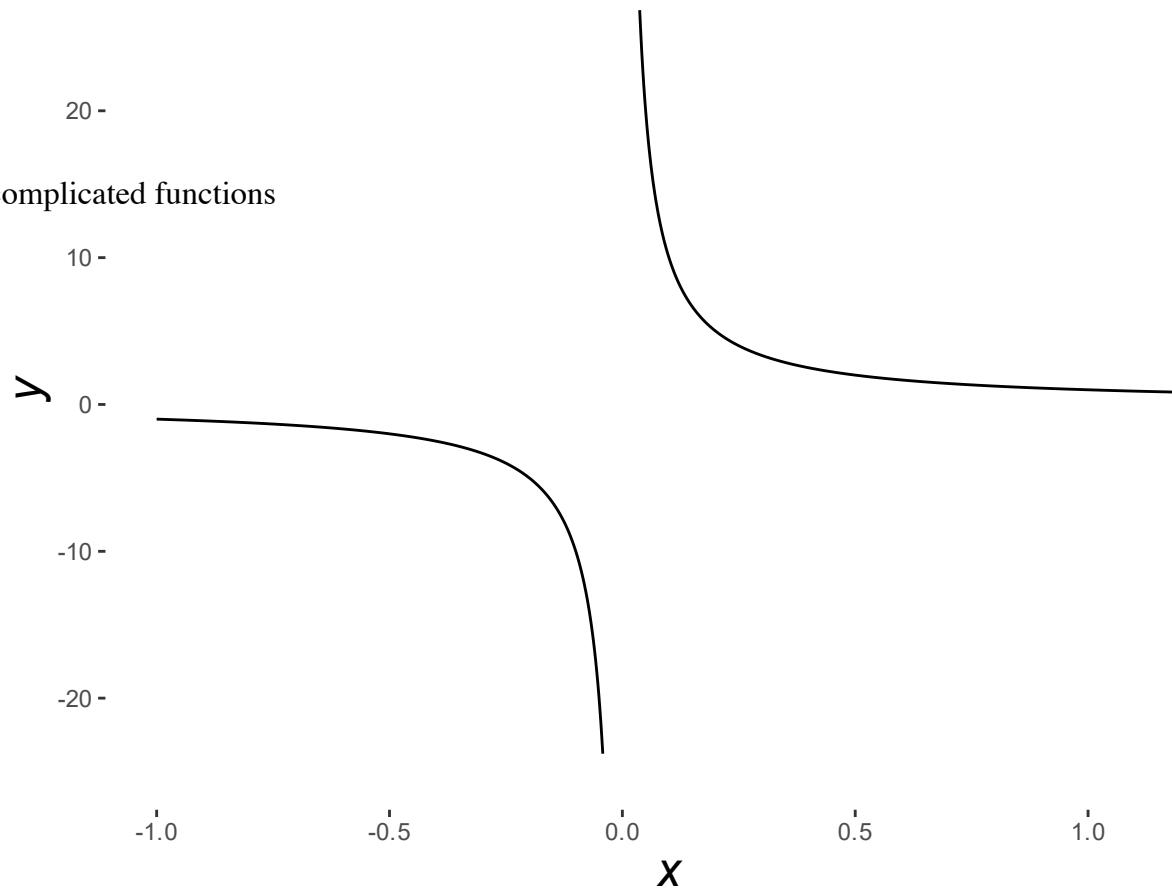
Polynomial have to be positive integers

Can have zero divisor as x, can go to infinity

If use polynomial for best fit and data is weird use rational > can create more complicated functions

- Polynomial functions are a subset of rational functions.
- Rational functions can be *discontinuous*: division by zero.

$$f(x) = \frac{1}{x}$$

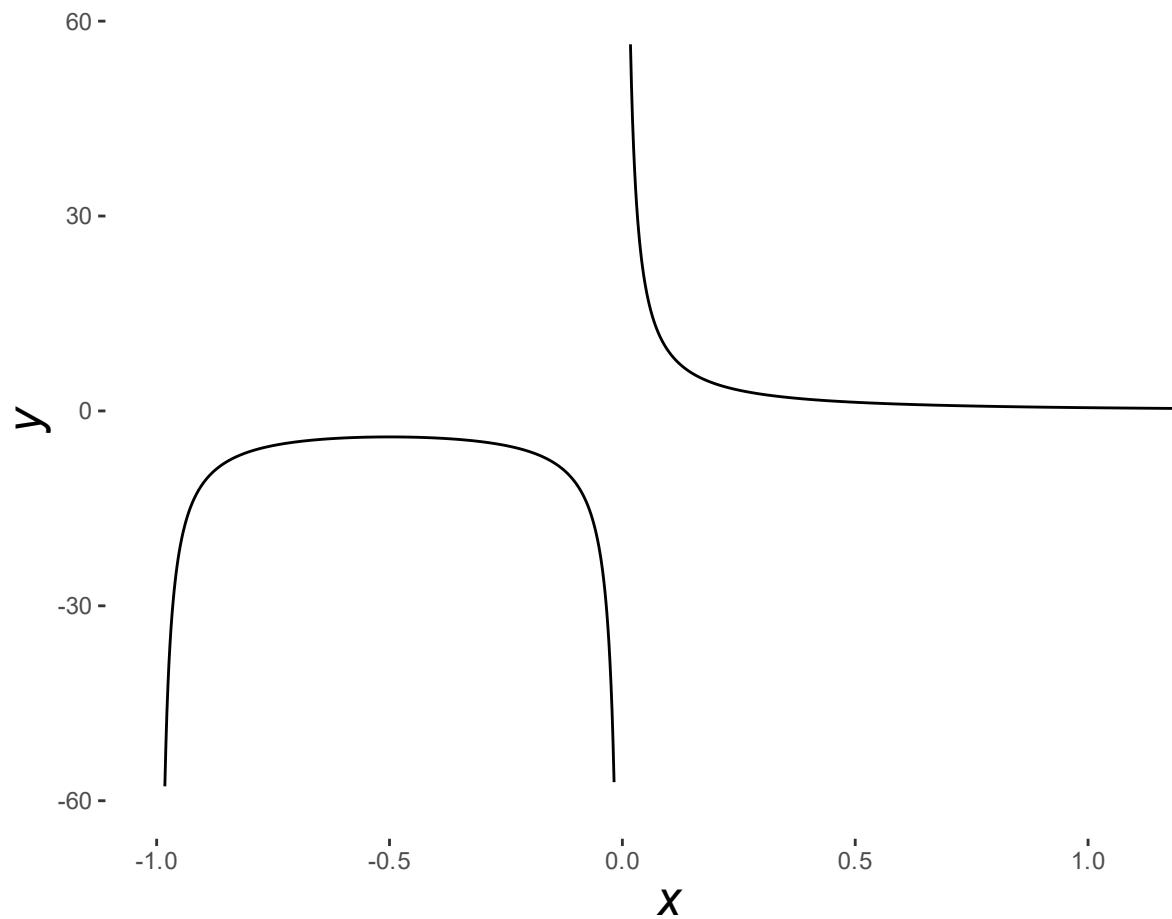


Rational functions

Rational functions are typically used in phenomenological models.

- Rational functions can emulate very complicated curves
- Tuning, improving normality of residuals, etc.
- Not used as often as polynomial or power law/fractional exponent functions

$$f(x) = \frac{1}{x + x^2}$$



Fractional or real number exponents

Functions in which the exponents can be expressed as fractions (rational numbers).

McGarigal calls these *power law functions*.

The square root function *is* a fractional exponent:

$$\sqrt{x} = x^{\frac{1}{2}}$$

Rational functions are typically used in phenomenological models.

- Often a result of *tuning* procedures like the Box-Cox transformation.

Fractional:

Polynomial and rational subsets of power functions

Can have non integer exponents

Have even more flexibility in functions to create

If have complicated curve structure use this

- There are so-called power-law distributions (like the Pareto), but we won't be talking much about these

Power vs. exponential functions

Which of these functions grows fastest?

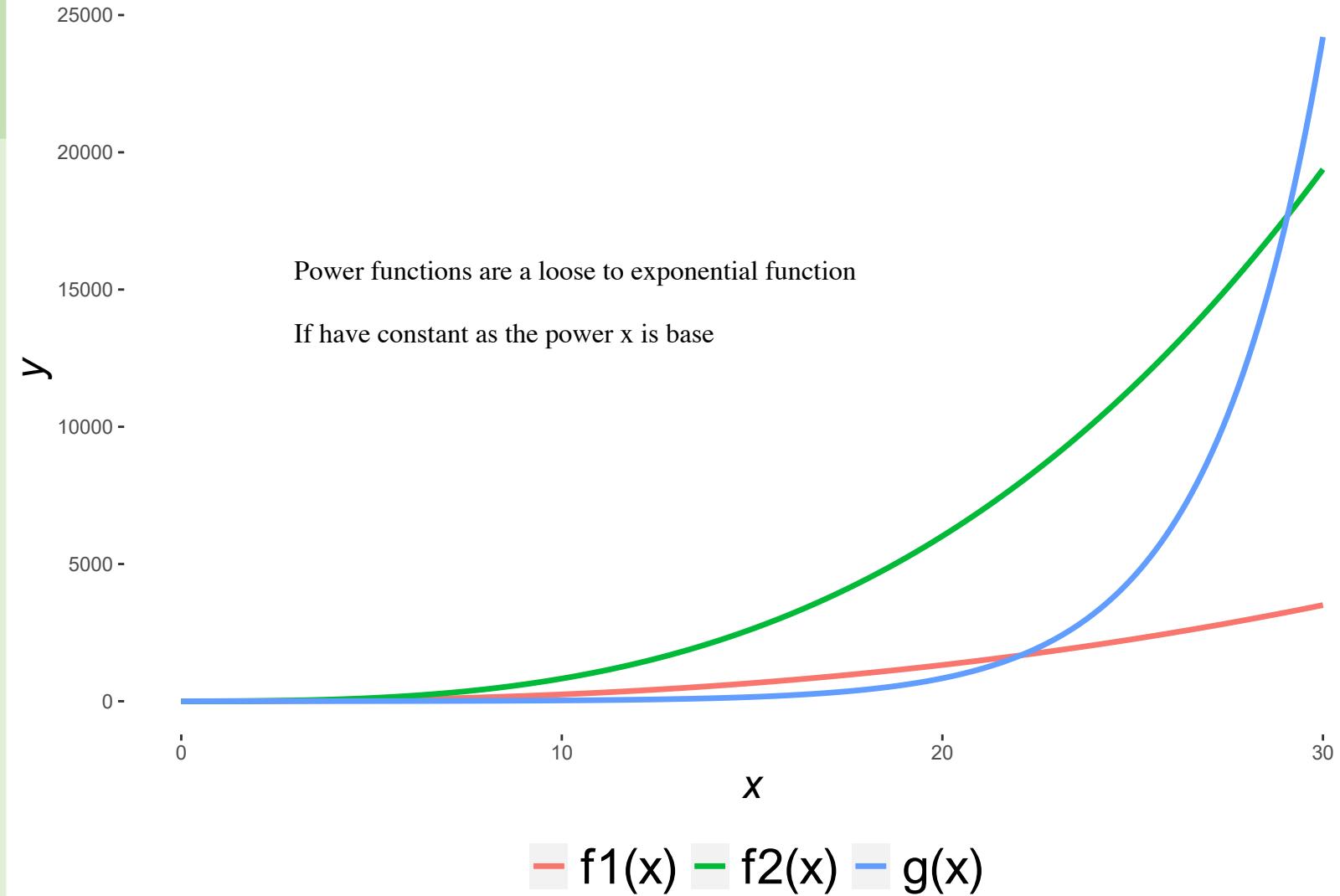
$$f1(x) = x^{2.4} - x^{0.5}$$

$$f2(x) = x^{2.9} + x^{1.5}$$

$$g(x) = 1.4^x - 0.1^x$$

- Exponential will always* win, eventually

*subject to terms and conditions



Power vs. exponential functions

In the **long term** exponentials always grow faster than any power (i.e. rational) function.

- A rational function may grow faster initially, but an exponential term always wins as x approaches infinity.
- An exponential beats any power. But the *gamma* function wins against an exponential...

Powers: the variable is the base; the power is a constant.

$$f(x) = x^{2.4} - x^{0.5}$$

Text

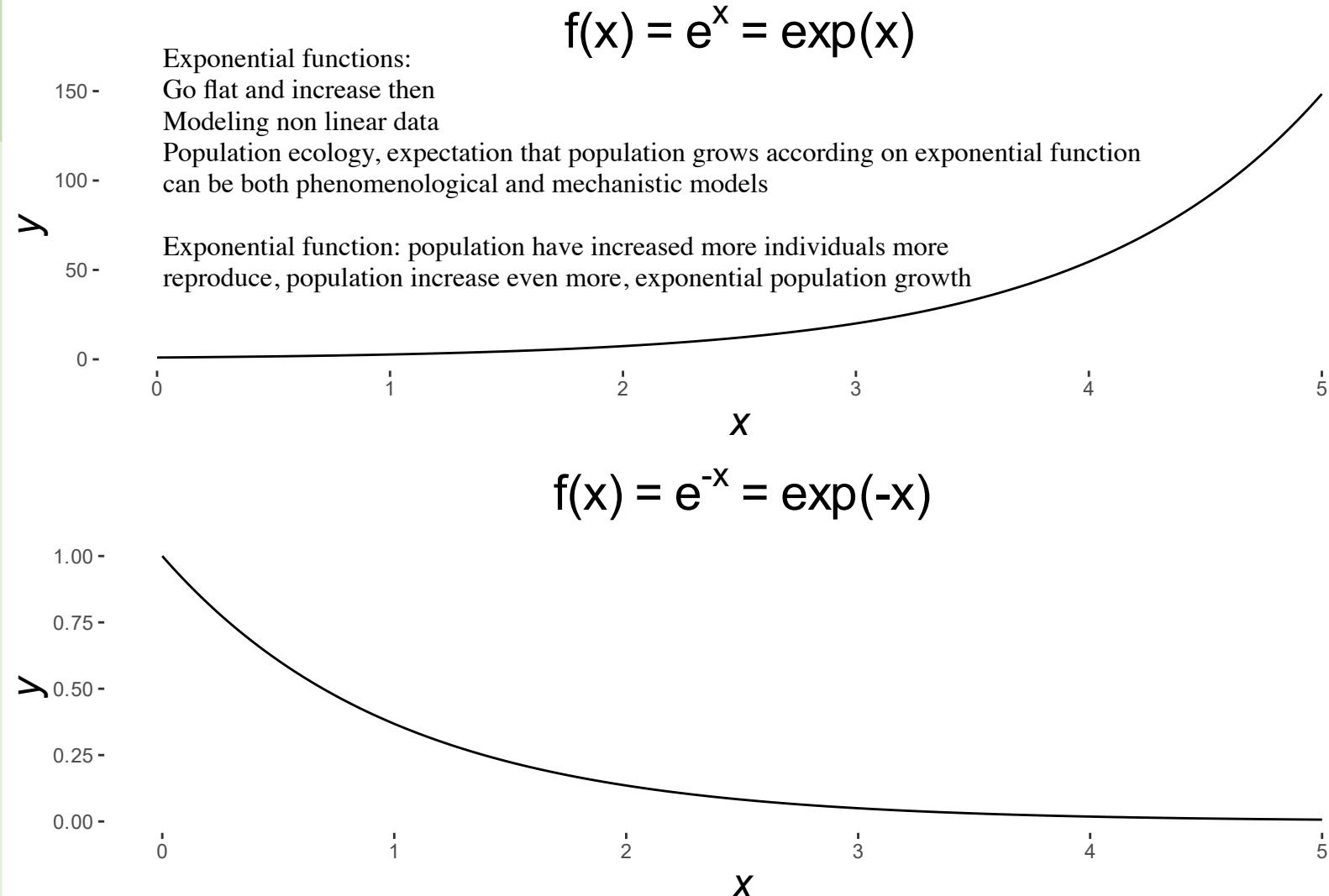
Exponentials: the variable is the exponent; the base is a constant.

$$g(x) = 2.4^x - 0.5^x$$

Exponential functions

Exponential functions have the variable as the exponent.

- When $x > 0$ the function is *monotonic increasing*.
- When $x < 0$ the function is *monotonic decreasing* and *asymptotic*.
- Any constant raised to the power of zero equals 1: $x^0 = 1$

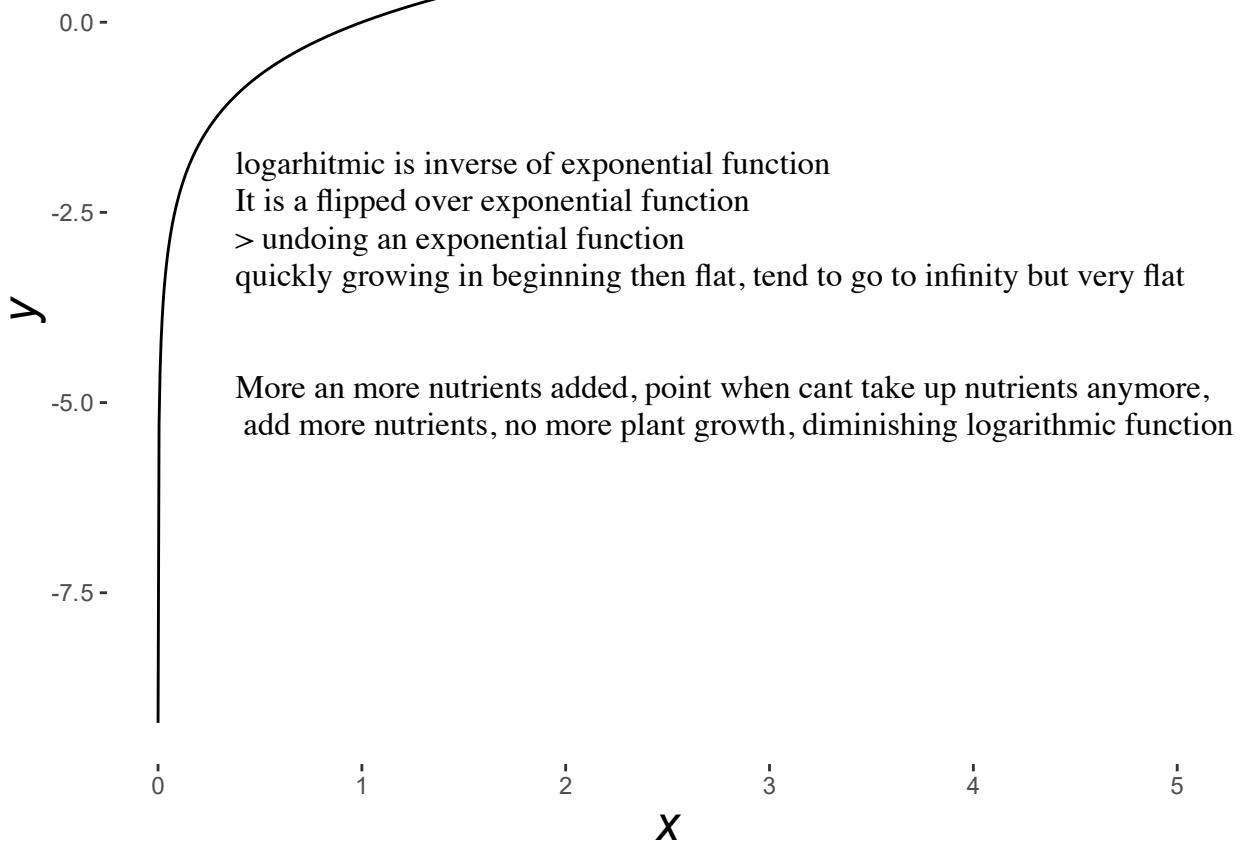


Logarithmic functions

Logarithmic functions are the inverse of exponential functions.

- Applying a logarithmic function *undoes* an exponential function.
- Logarithmic functions are slow-growing, but *not asymptotic*.

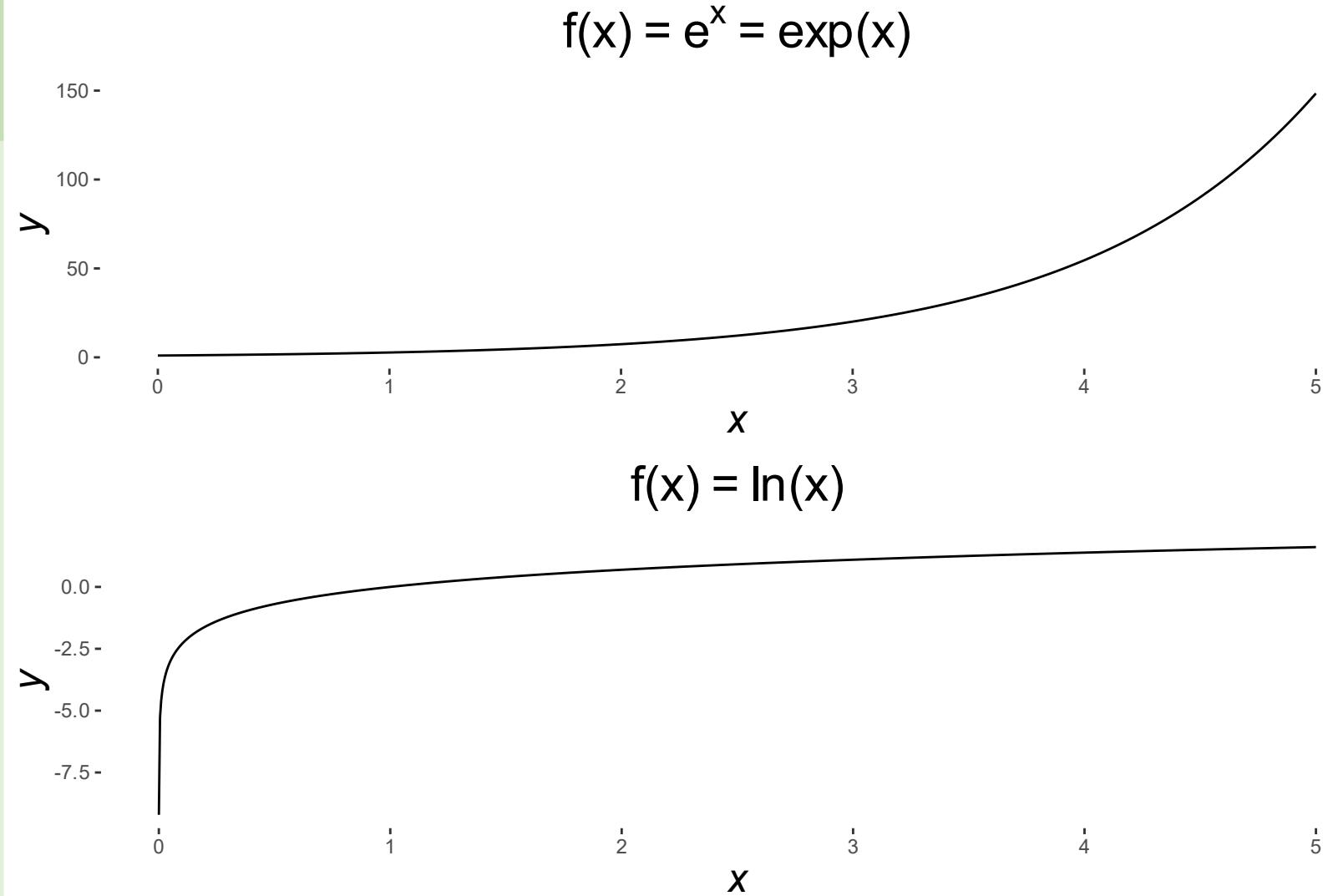
$$f(x) = \ln(x)$$



Exponential and Logarithmic functions

Mechanistic Interpretations

- Exp: *feedback processes, exponential growth, divergent*
- Log: *diminishing returns, useful for dealing with very large number, linearizing, variance stabilizing*



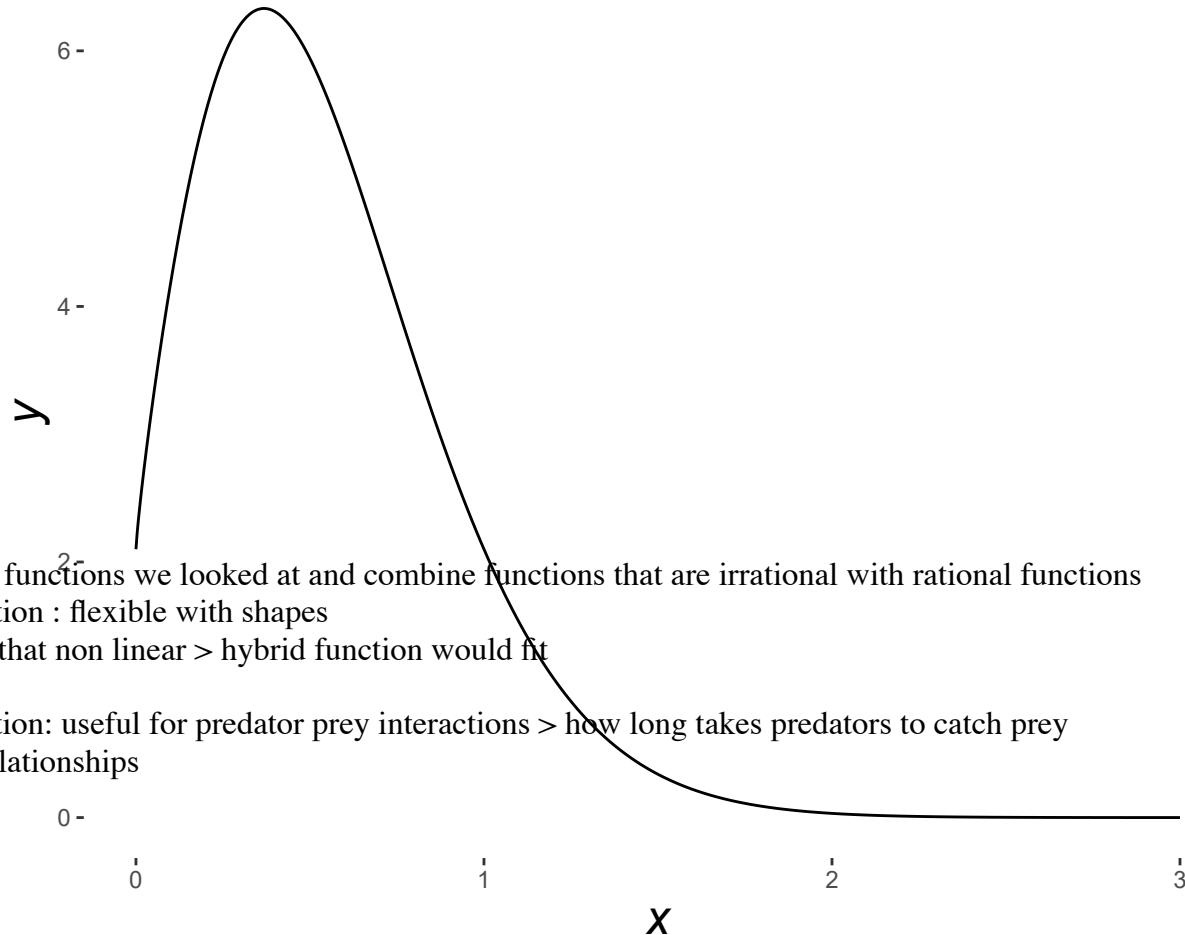
Hybrid functions

Just like the name says,
they are mixtures of
different function types.

- Often have a theoretical basis: they can be *mechanistic*.

The Ricker function

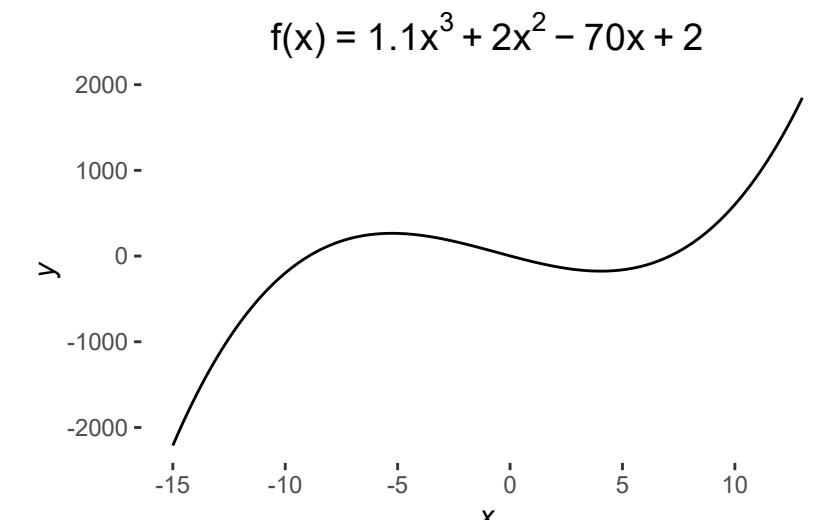
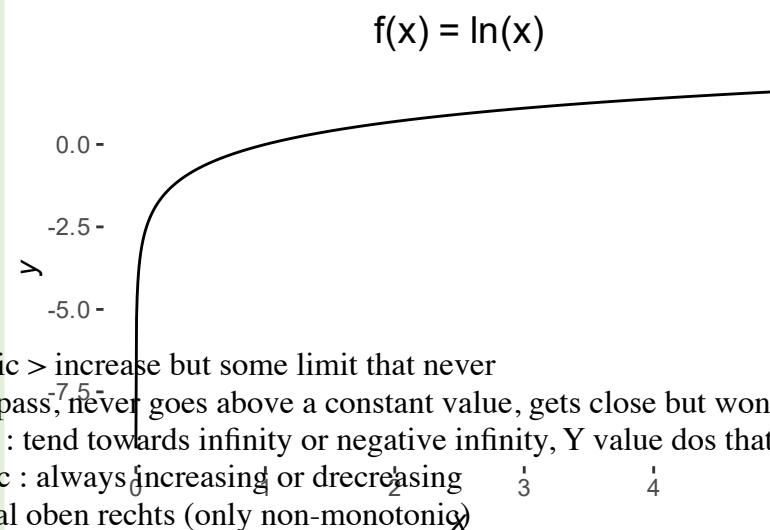
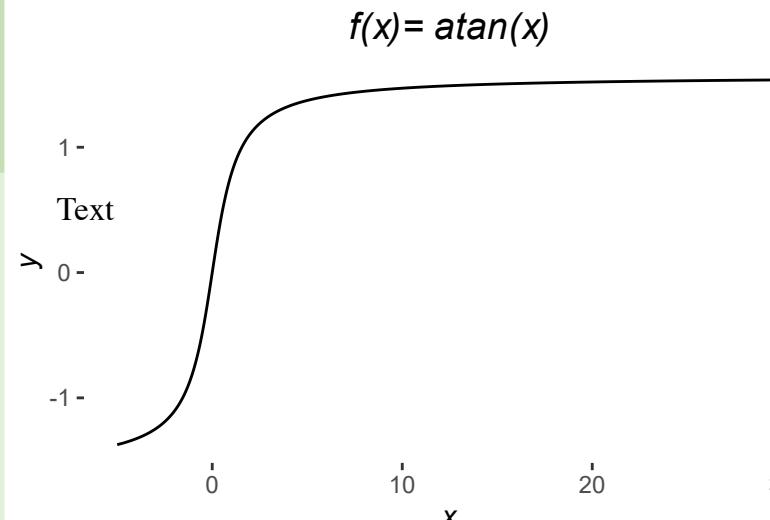
Ricker Function: $f(x) = 2.1x^{-3x}$



Graphical intuition

Function Terminology

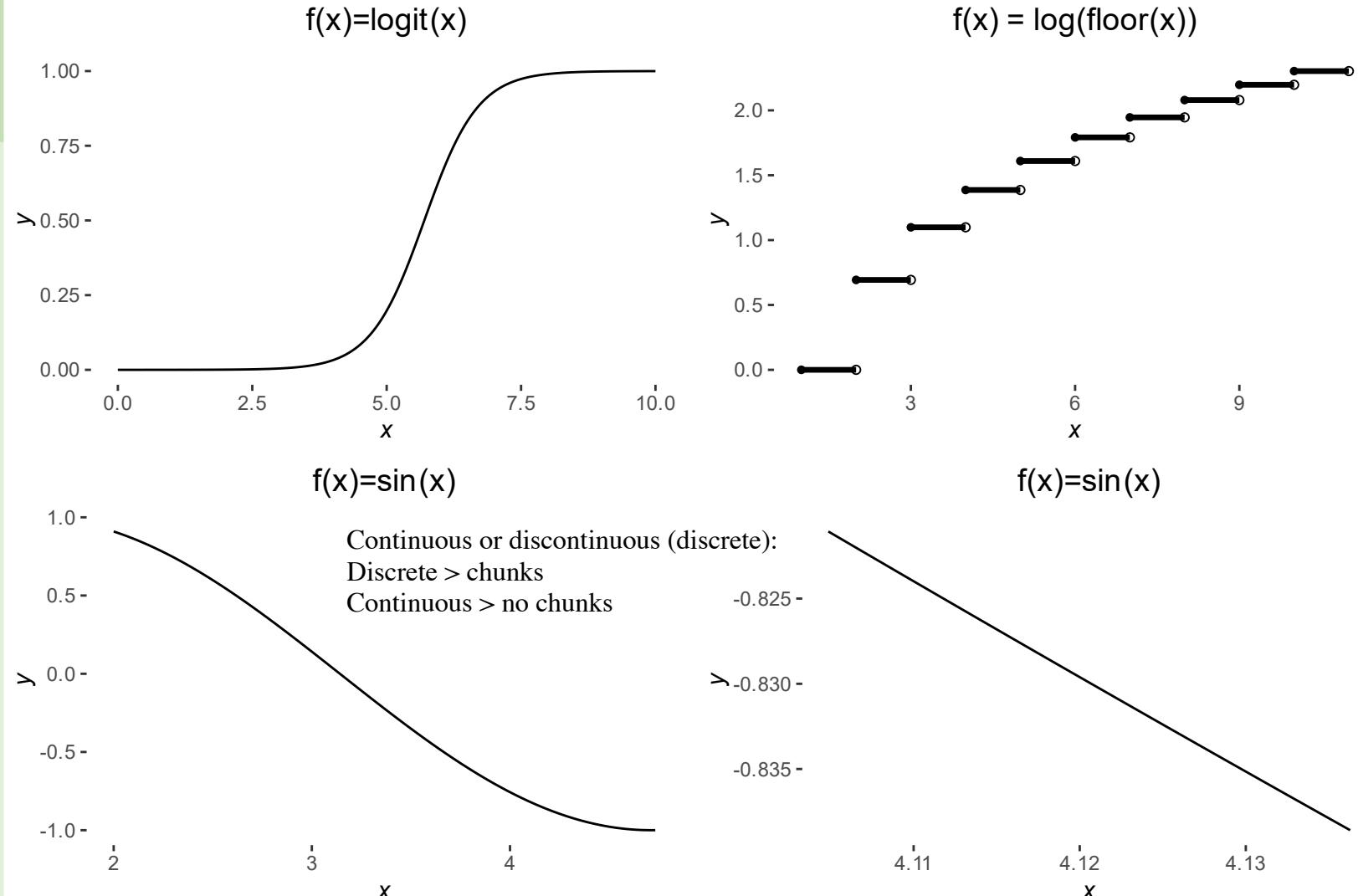
- Asymptotic: tends toward a value
- Divergent: tends toward infinity or negative infinity
- Monotonic: always increasing or always decreasing



Graphical intuition

Function Terminology

- Continuous: no breaks or jumps
- Local linearity: *most* functions resemble linear functions if you zoom in close enough.
 - This is closely related to **differentiability** in calculus.



In-Class Reading Data Files

Probability Distributions 1

General Concepts

Key probability terms and concepts

- Inference with the dual model paradigm
- What is a distribution?
- Event, domain, sample space
 - domain and sample space interchangeable
- Key probability theory results
 - law of total probability
 - independent events



Stepping back: What do we need to do inference?

We need a model: a *dual* model!



Why do we want to do inference?

- We want to go beyond descriptive statistics.
- We want to learn something about a larger population from a sample.
- We want to estimate population parameters from sample statistics.
- We want to create a statistical model for understanding and/or prediction

Take statistics to make inferences about ecological parameters

Stepping back: What do we need to do inference?

We need the Dual Model Paradigm to do inferential statistics.

Need Dual Model paradigm to make inferences

> Need deterministic model: average

And also have noise > here comes probability distribution > characterises stochastic model

Deterministic model by function, Stochastic by probability distribution

If seen something unusual: stochastic version helps to determine

- We need the *deterministic* model of the means to about the *average* or *expected* behavior.
- We need the *stochastic* model to know about the variation.
- We need the *stochastic* model to know if an observation is *unusual*



Stepping back: What is inference?

Deterministic system has randomness determined by a linear model > LINEAR FUNCTION IS DETERMINISTIC MODEL
Don't see how much variation to expect > therefore need stochastic model

For our purposes: inference is a way to learn something about a larger *population* from the properties of a *sample*.

More formally: Inference is estimating population *parameters* from sample *statistics*.

- We use the *deterministic* model to calculate model parameter estimates.
- We use the *stochastic* model to quantify *confidence* and *significance*.

Inference: why do we need distributions?

Couldn't we just use our deterministic model to make predictions?

- Sure, but without a stochastic model we can't quantify the uncertainty in our guesses.
- Relatively few systems are completely, or even mostly described by a deterministic model.
 - Planetary orbits
 - Chaotic systems governed by deterministic functions: sadly, we won't get to talk about these.
 - Logistic population growth
 - Lorentz equation

Probability Distributions

- Help us understand the 'noise' part of the system.
- Help us quantify and understand uncertainty.
- Theoretical Distributions
 - There are hundreds of named, parametric distributions
 - Defined by mathematical functions
 - Describe Stochastic processes
- Empirical Distributions
 - Calculated from data

Empirical > characterises what randomness in population looks like by calculating the spread/properties of noise
Theoretical distribution we prefer to use!

What is a distribution?

Remember that words often have specific meanings in statistics:

- What do I mean by *likelihood*?
- What do I mean by *event*?

A distribution is a map from events to measures of likelihood

- Why would we want such a map?
- What do I mean by likelihood?
- We'll talk about probability theory later.

when think about probability distribution > map that takes events and likelihood

Likelihood > how common particular events are > stochastic part of system, likelihood of all possible events, out of 100 each has 1% chance of happening, one event has to happen, so total probability is 100%

Theoretical distribution > have equations, that help with distribution functions that help determine likelihood, helpful to compare null and alternative hypotheses

Parametric and Empirical Distributions

Parametric distributions are defined by mathematical *functions*

- The functions have one or more *parameters* that define how probabilities are allocated to events.
 - What are the parameters of the Normal distribution?
 - We often want to estimate the parameters from samples.

Empirical distributions are computed from *observations*.

- There is no analytical function, but we can compare empirical distributions to parametric distributions.
- Useful for comparing null and alternative hypotheses

Probability Distribution Functions

Probability density function>always positive, non/monotonic

The map of events to probabilities are defined by:

- **Probability Density Functions** for continuous distributions
- **Probability Mass Functions** for discrete distributions.
- The values of PDFs and PMFs are always non-negative, by the definition of probability.

Text

Two other types of functions are used to describe distributions:

- **cumulative functions**
- **quantile functions.**

Described by either density function or probability mass functions
Main types of propability functions > density and mass function
Two other:
Cumulative and quantile

Density or Mass Function: PDFs & PMFs



Value x, likelihood is related to y value of function
Discrete distribution probability of x happening is y value
Height of density > gives relative measure of likelihood > is continuous

Probability density is the y-value of the probability density curve for a given value of x.

- You can think of it as the height of a curve
- For *continuous distributions*, it is *not* equal to the probability of observing a particular value of x.

Cumulative Probability Functions: CDFs & CMFs

Probability Density is the **height of the density curve**.

- Provides a measure of likelihood of an event
- Measure is relative for continuous; measure is the probability for discrete.

Cumulative density is the **accumulated area under the density curve** to the left of x .

Cumulative distribution function > accumulated

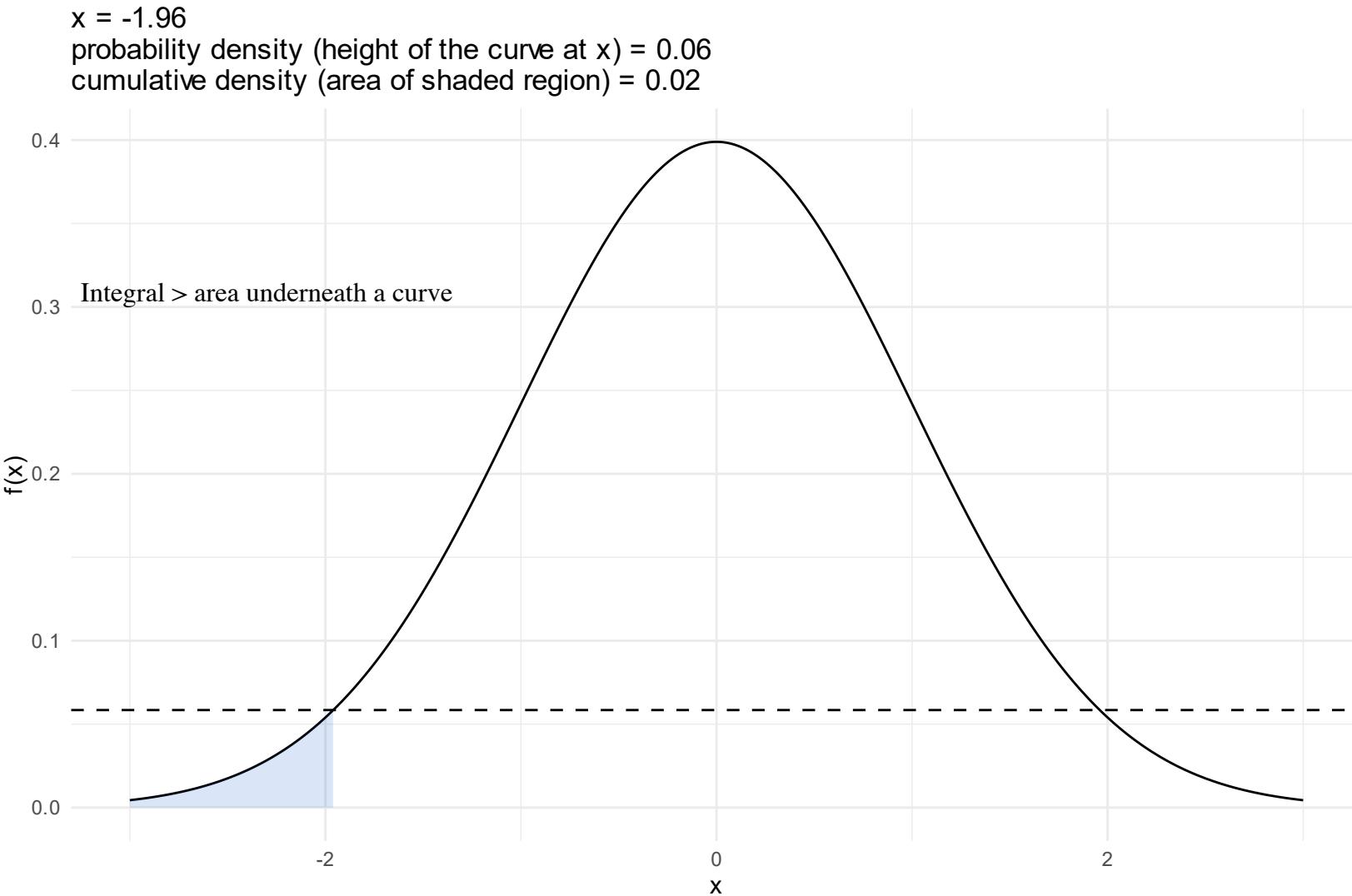
- It's an integral!
- It is the probability of observing a value equal to or less than x .

Probability Distribution Functions: Graphical Intuition

Demonstration of PDF and CDF using the Normal distribution.

Remember:

- *Density* = height of the curve at x .
- *Cumulative Density* = area under the curve, to the left of x

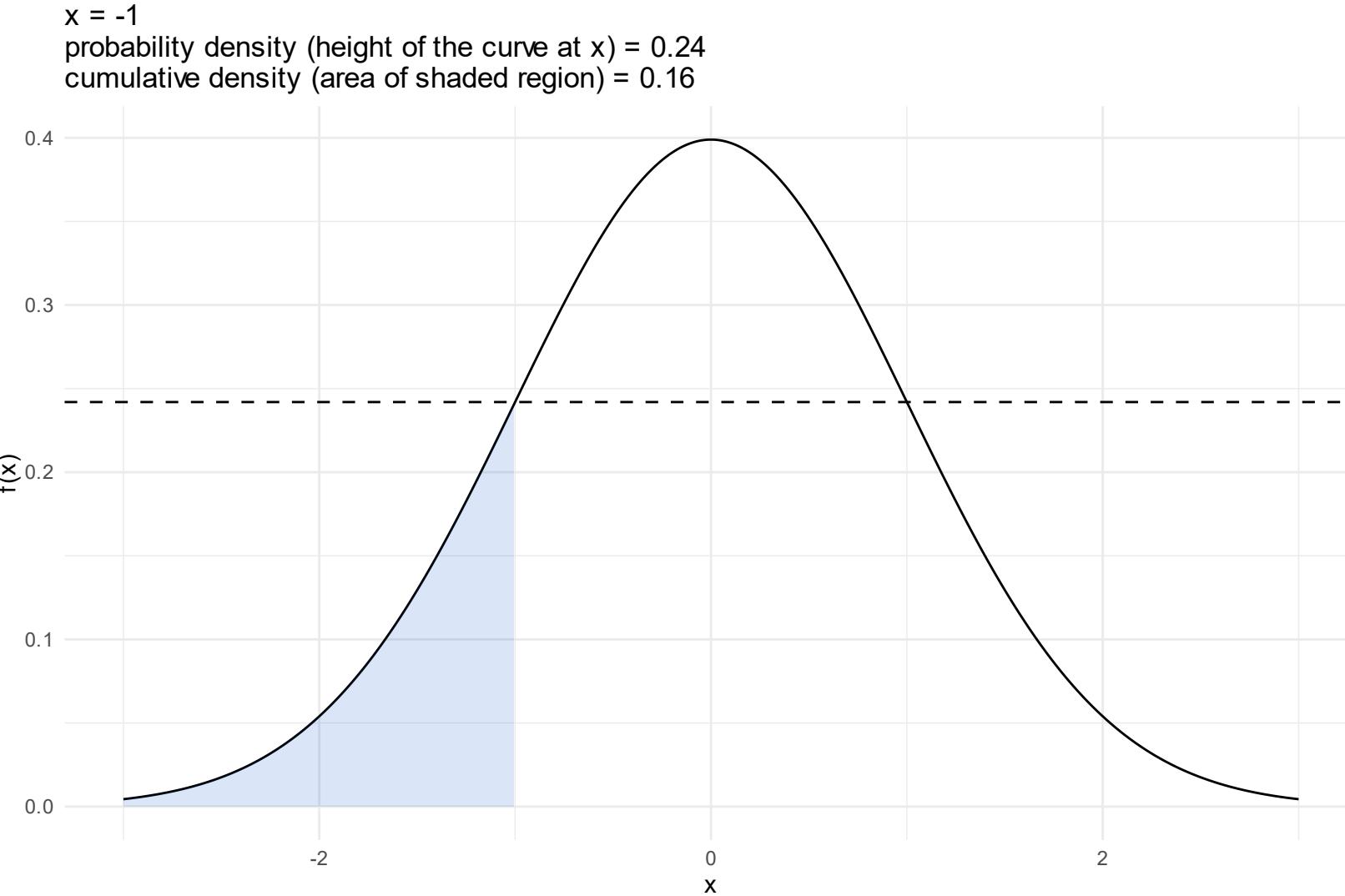


Probability Distribution Functions: Graphical Intuition

Demonstration of PDF and CDF using the Normal distribution.

Remember:

- *Density* = height of the curve at x .
- *Cumulative Density* = area under the curve, to the left of x

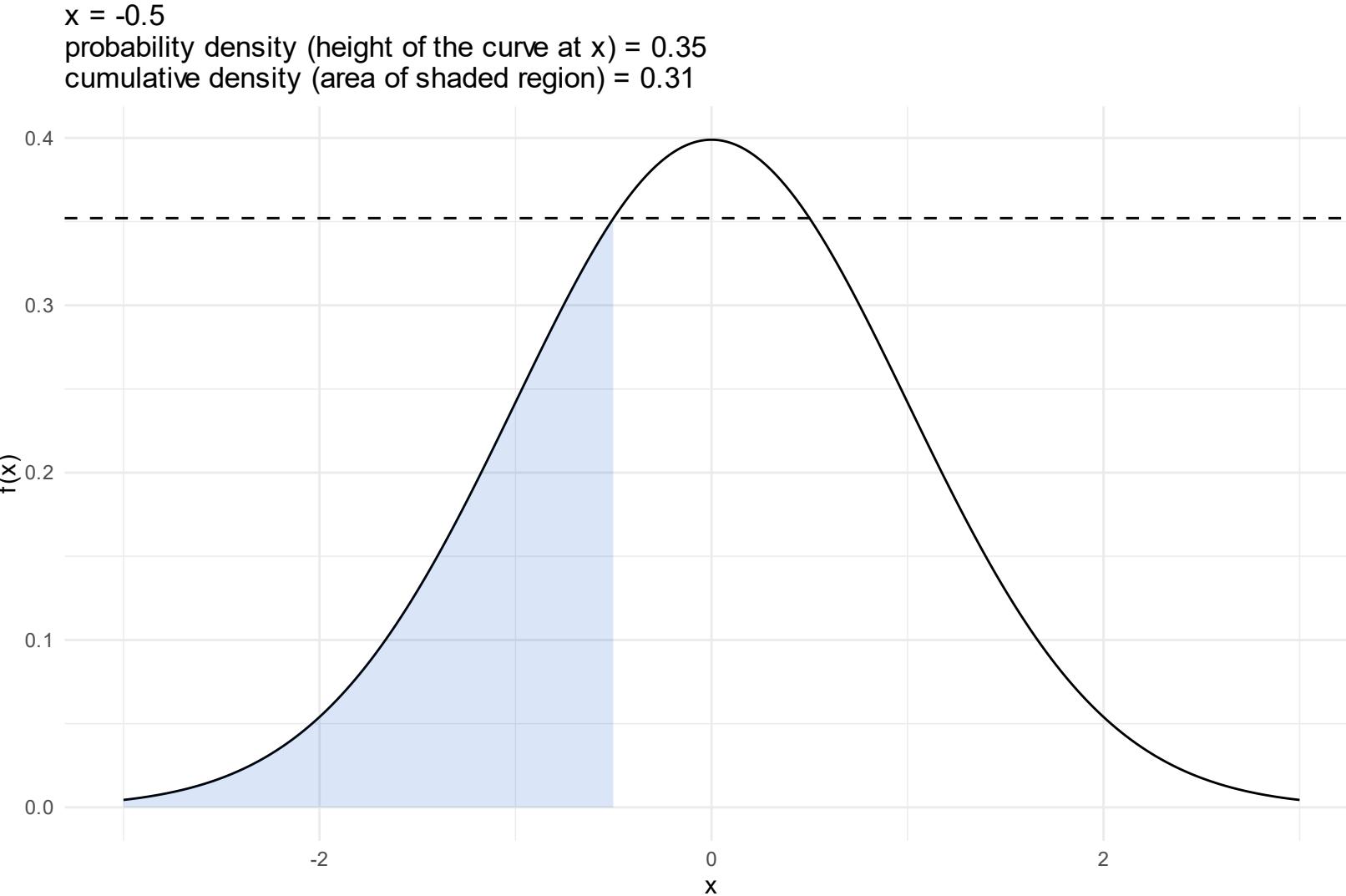


Probability Distribution Functions: Graphical Intuition

Demonstration of PDF and CDF using the Normal distribution.

Remember:

- *Density* = height of the curve at x .
- *Cumulative Density* = area under the curve, to the left of x

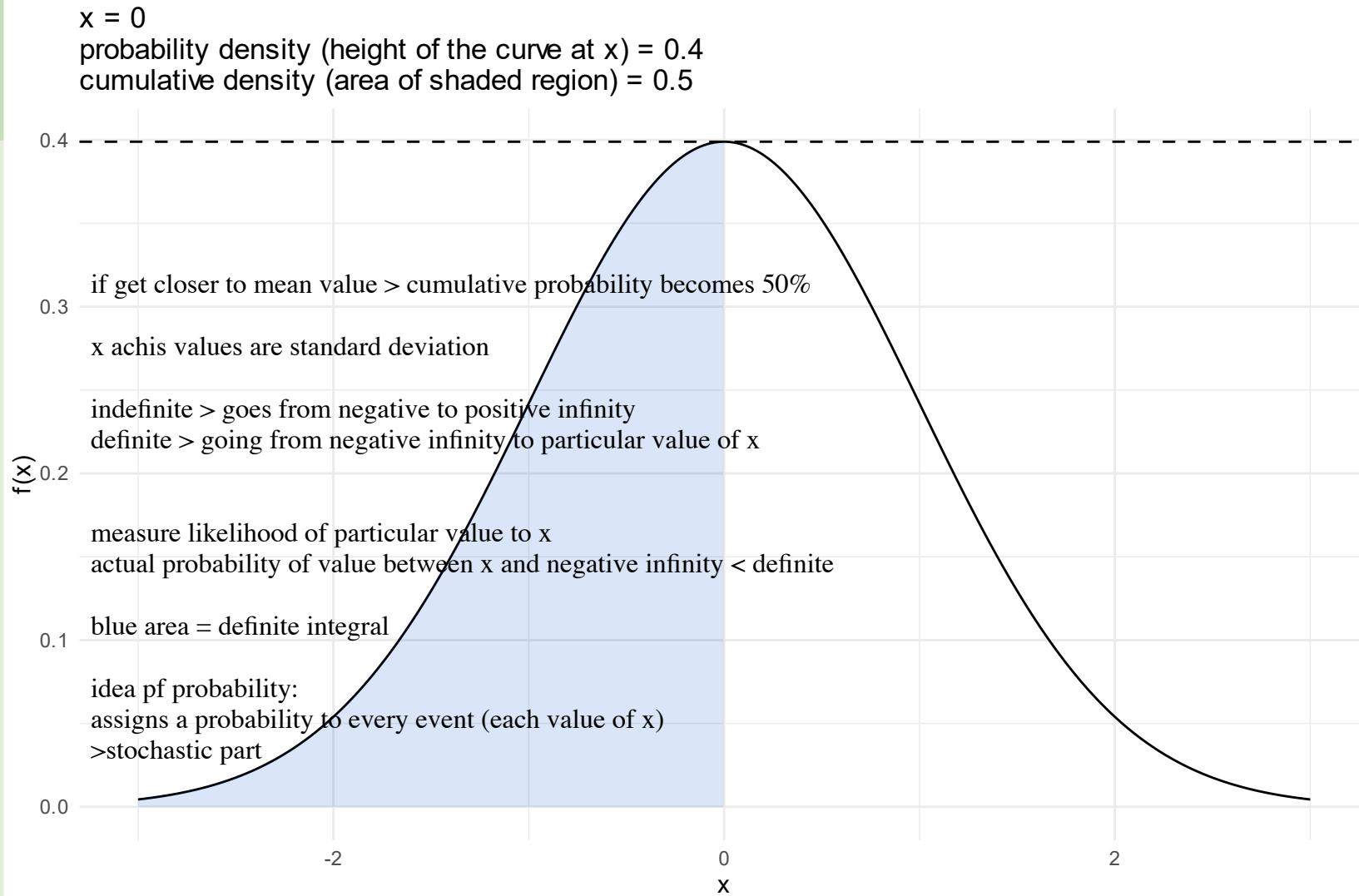


Probability Distribution Functions: Graphical Intuition

Demonstration of PDF and CDF using the Normal distribution.

Remember:

- *Density* = height of the curve at x .
- *Cumulative Density* = area under the curve, to the left of x

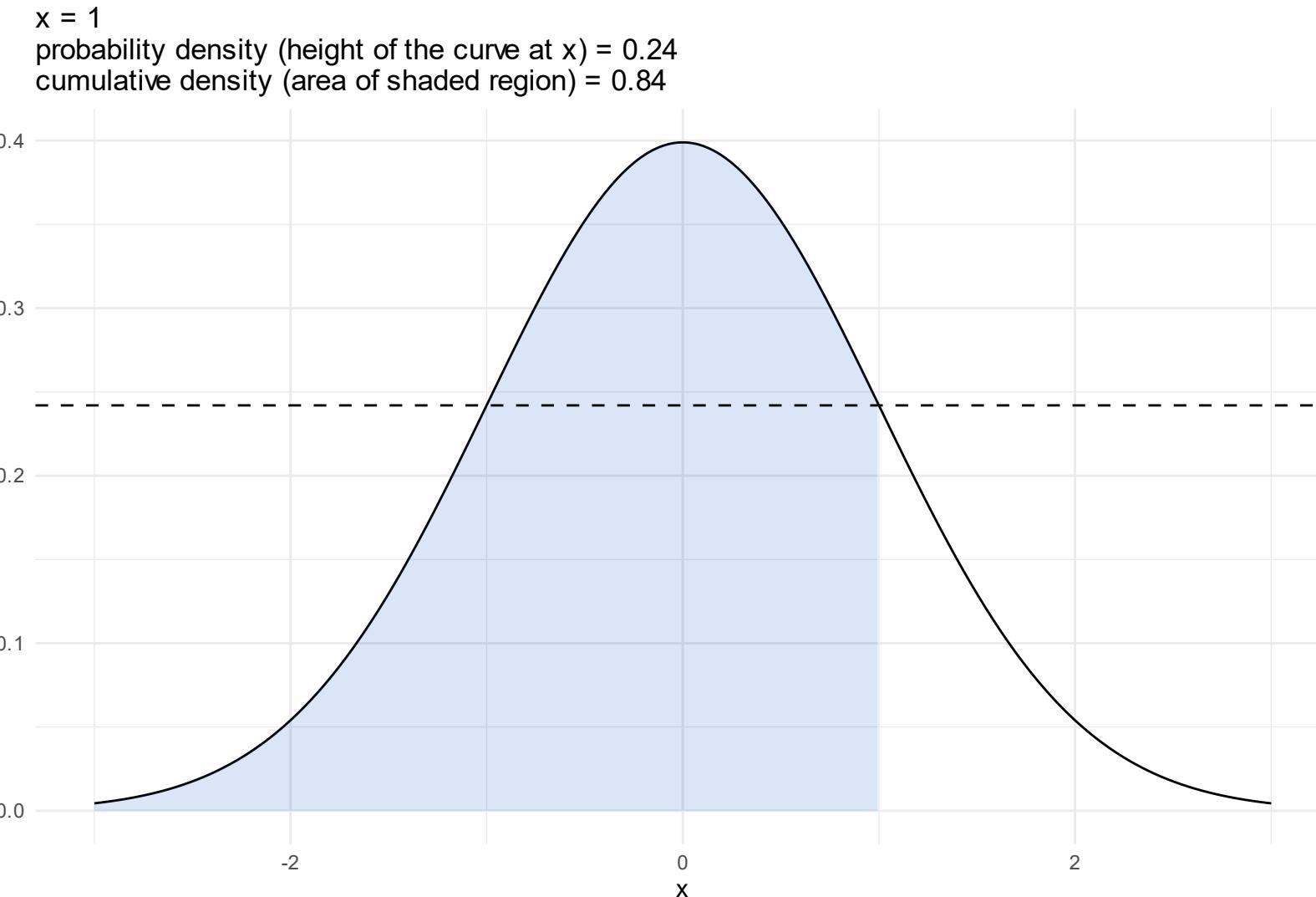


Probability Distribution Functions: Graphical Intuition

Demonstration of PDF and CDF using the Normal distribution.

Remember:

- *Density* = height of the curve at x .
- *Cumulative Density* = area under the curve, to the left of x

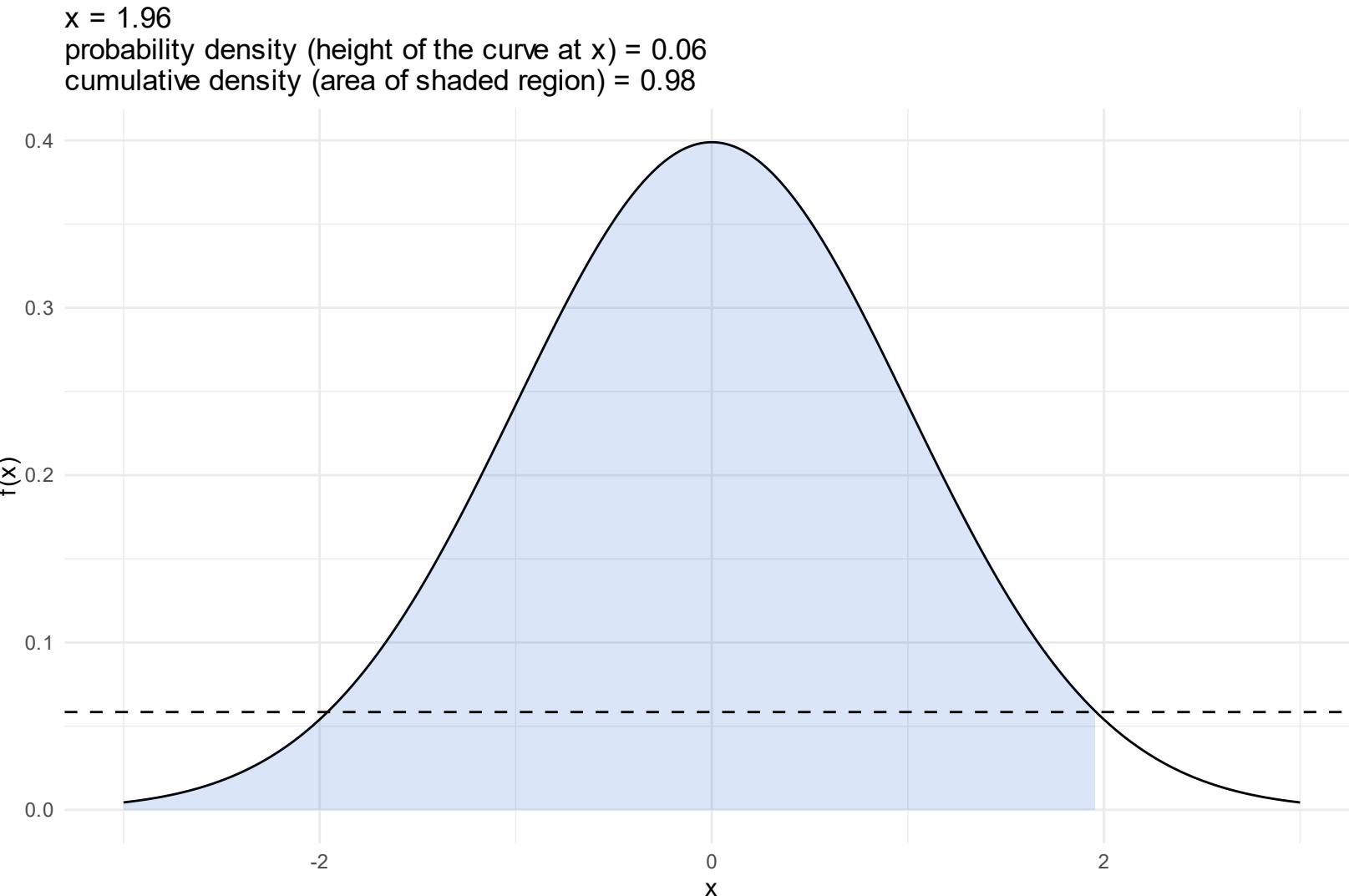


Probability Distribution Functions: Graphical Intuition

Demonstration of PDF and CDF using the Normal distribution.

Remember:

- *Density* = height of the curve at x .
- *Cumulative Density* = area under the curve, to the left of x

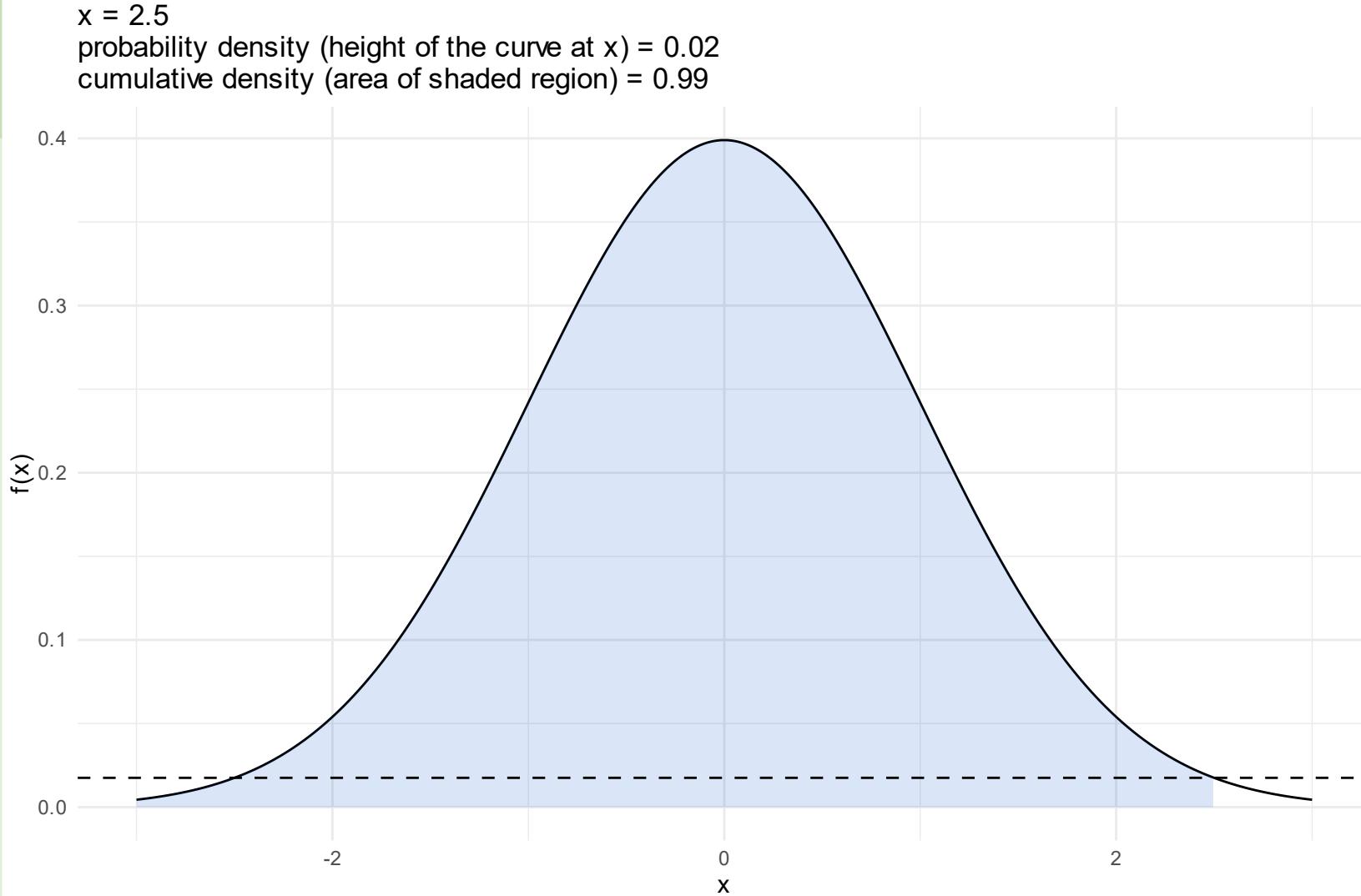


Probability Distribution Functions: Graphical Intuition

Demonstration of PDF and CDF using the Normal distribution.

Remember:

- *Density* = height of the curve at x .
- *Cumulative Density* = area under the curve, to the left of x



Recap of essentials:

Distributions

1. They assign a *probability* to every *event* in a *sample space*.
2. We can use them as the *stochastic model* in the dual model paradigm.

Probability essentials

1. Probabilities are non-negative
2. Law of Total Probability: Probabilities of all events in sample space sum to 1.0
3. Independent events: joint probability is product of individual probabilities

We'll continue to build our intuition about Probability Distributions