Analysis of Environmental Data

Data Exploration, Functions, and Associations

Michael France Nelson

Eco 602 – University of Massachusetts, Amherst – Fall 2021

Michael France Nelson

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
• Can contain any kind of R object • 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	• 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.

Absolute and Relative Paths - Review

- Absolute paths are rooted at your storage device's root folder.
  - On windows this usually C:\
  - On Mac it is usually called MacHD or something similar.
- Relative path is like a set of turns from a starting point to a destination
  - Your course Rproject folder might be a great place to start!
- Absolute paths are machine-specific
- Relative paths are more flexible.
  - As long as you have your course folders organized the same way, you can exchange code with colleagues in this course.

The here() function and your RProject folder

- The here() function always starts at your Rproject folder
- You should never use setwd()
  - It's important to know about working directories, but you should avoid setting them manually.
  - setwd() leads to fragile code.
- Code smell: there's something fishy in your code.
  - Code smell indicates potential problems with your code
  - Code smell doesn't necessarily mean your code won't run, but it can be easily broken; it may be difficult to share with others.

Announcement: Grading

- Please double check your submitted assignments.
- We're still working out the Moodle grading kinks.
- Verify that you can see the feedback.
- Let us know if you think there is an error in grading!
- Keep finding those typos! There were some good ones in lab yesterday!

Announcement: Data Exploration 1 and 2 Are Swapped

- Mea culpa, I switched the names of the assignments.
- Today we'll do Data Exploration 1.
  - Don't worry, it's a lot simpler than D.E. 2

Schedule: Week 4

- Tuesday
  - Continue this slide deck
  - Reading data files in-class exercise.
  - Finish data exploration from Thursday (as needed)

- Thursday
  - Finish this slide deck
  - Data exploration in-class part 2 (the real part 2)

Announcements: Sep 21

- Mike: make sure your screen is shared and you restart zoom recording!
- Today is our first Zoom poll!
  - Open the course zoom channel – the link is in Moodle
- Cue up today's in-class activity
- Data exploration/deterministic functions assignment now due the 26th
- Assignments are due Sundays at 11:59PM
  - Some assignments had the wrong due time – these are fixed.
  - If you submitted your assignment on Sunday, it is on time.

Announcements: Sep 23

- Can you believe it's already September 23rd?
- Today we'll do the real
- Correct version of the week 4 reading assignments are posted.
  - They are simply re-wordings of the draft questions.
  - Thanks to Bonnie for the reminder!
- Read the Library of Babel for discussion next week.
  - It's a fun read, it encapsulates a lot of probability theory in an interesting way.
  - We'll use insight from the reading for next week's in-class activity!
- In the other readings, pay special attention to the discrete probability and discrete probability distribution materials.

Announcements: Sep 23

- Updated final version of deck 3 pdf is on GitHub – You should re-download
- Today's Agenda:
  - Finish this deck!
  - In-Class data exploration 2. The real 2, the one with file import!
- Items due Sunday
  - Week 4 reading questions
  - Data exploration/deterministic functions
  - Three in-class group assignments, in case you haven't already submitted:
    - Data exploration 1 and 2
    - In-class data file exercise

What's In This Deck?

Slides	Selected Key Take-Home Concepts
• Functions, variables, constants • Formulae and notation • Classes of functions • Types of Plots	• 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.

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:

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

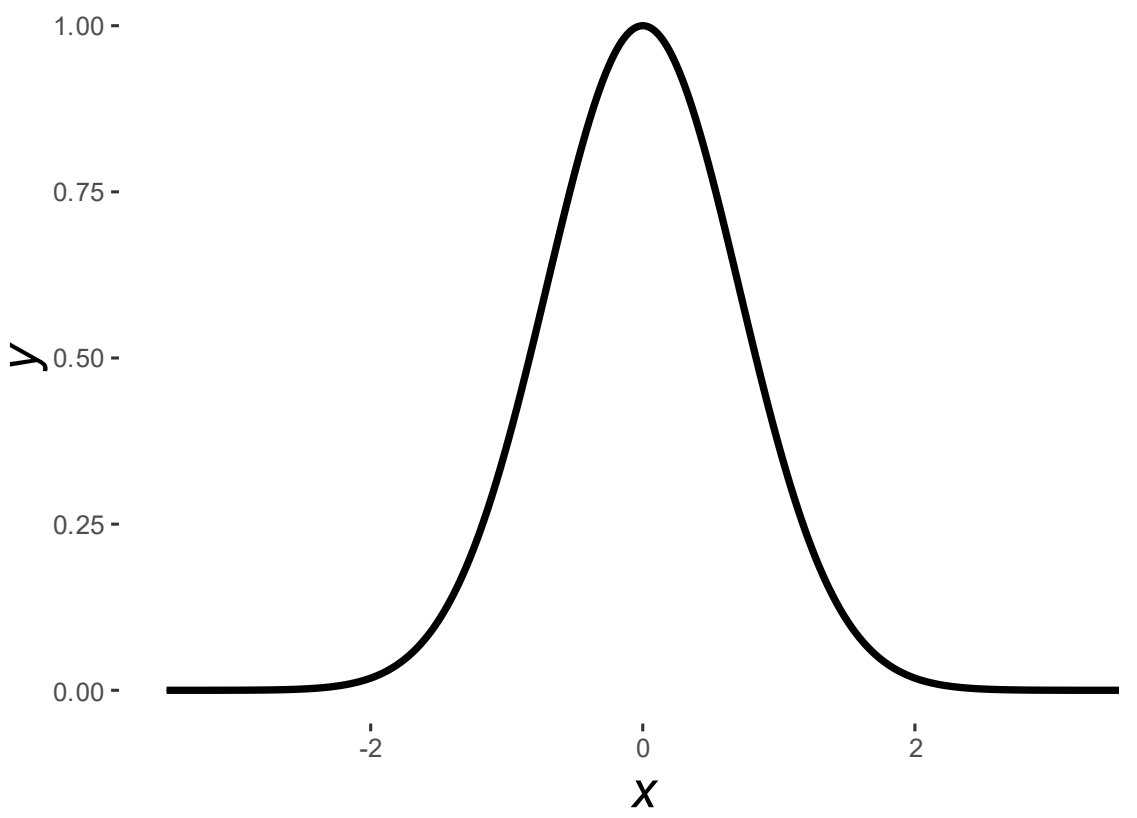
That is considerably simpler to understand than the original monstrosity!

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

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 determinstic models:

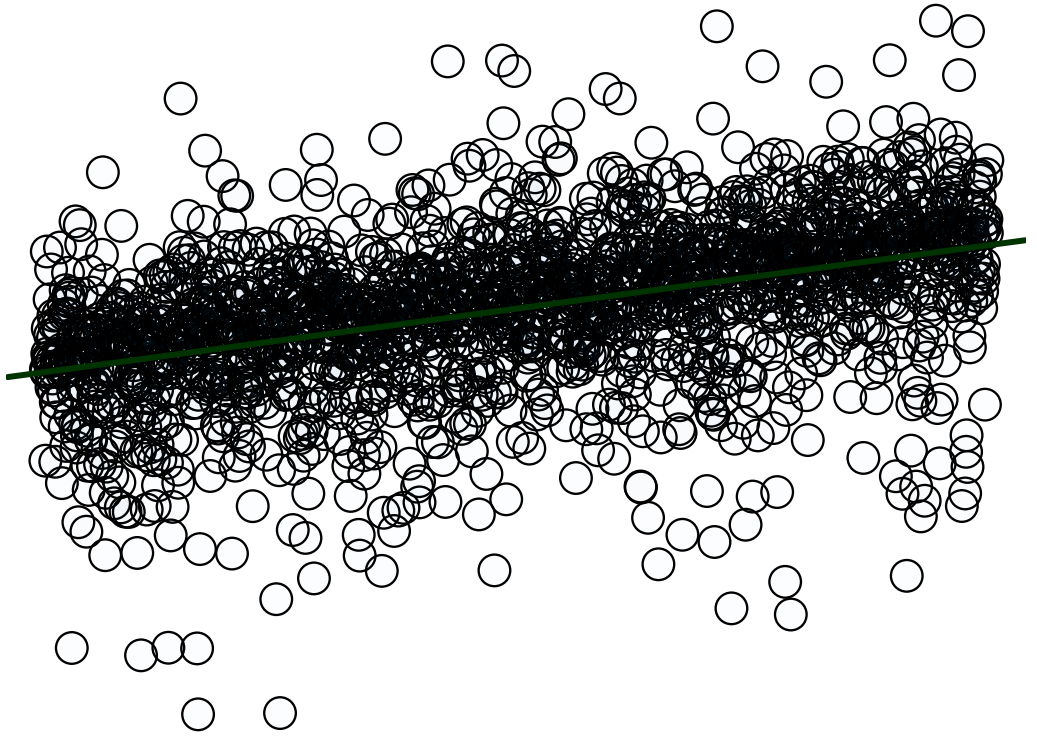
- linear
- polynomial, rational, and power
- exponential (and logarithmic)
- periodic
- combination functions

Linear functions

Linear functions have variables raised to a power of 1.

- Can be one or more variable variable:
  - $y = mx + b$
  - $y = m_1x_1 + m_2x_2 + \ldots + m_nx_n + b$
- Statistical literature likes to use alphas, and betas
  - $y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \ldots + \beta_nx_{ni} + \epsilon$
- Key features:
  - The *variables*, i.e. the x_i, are first-degree.
  - Each *variable* is multiplied by a *parameter*, the β_i

A linear model is always a great place to start.



Linear models: interpretation

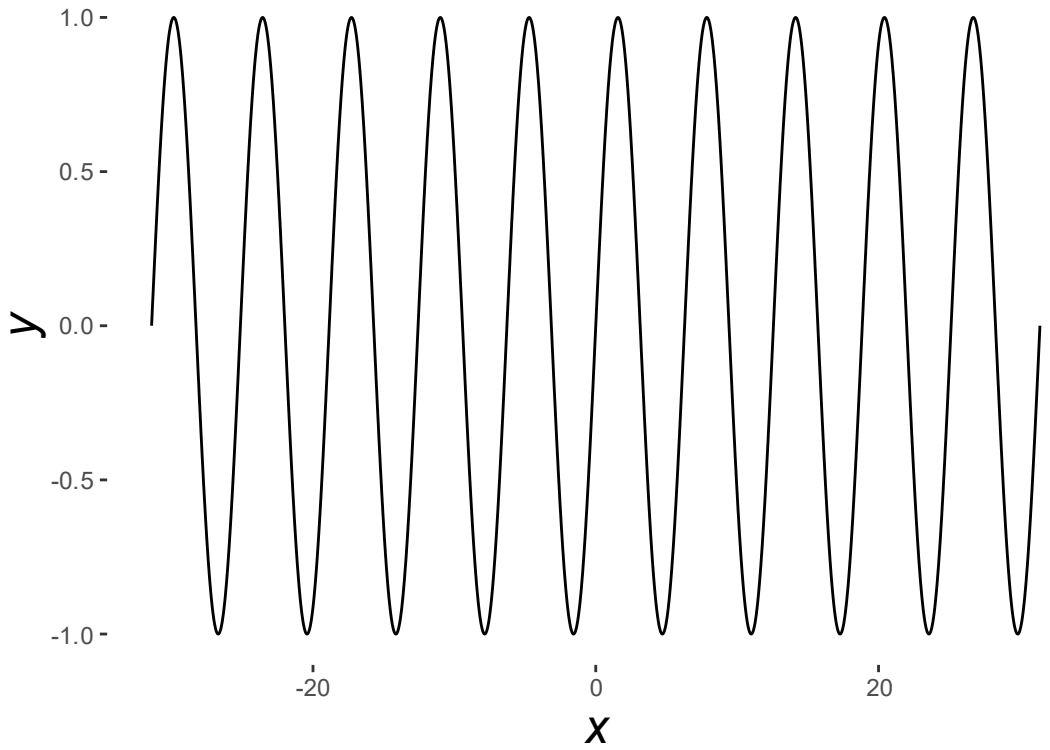
A linear model describes a *constant rate of change*

- 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.

Is the constant rate of change reasonable?

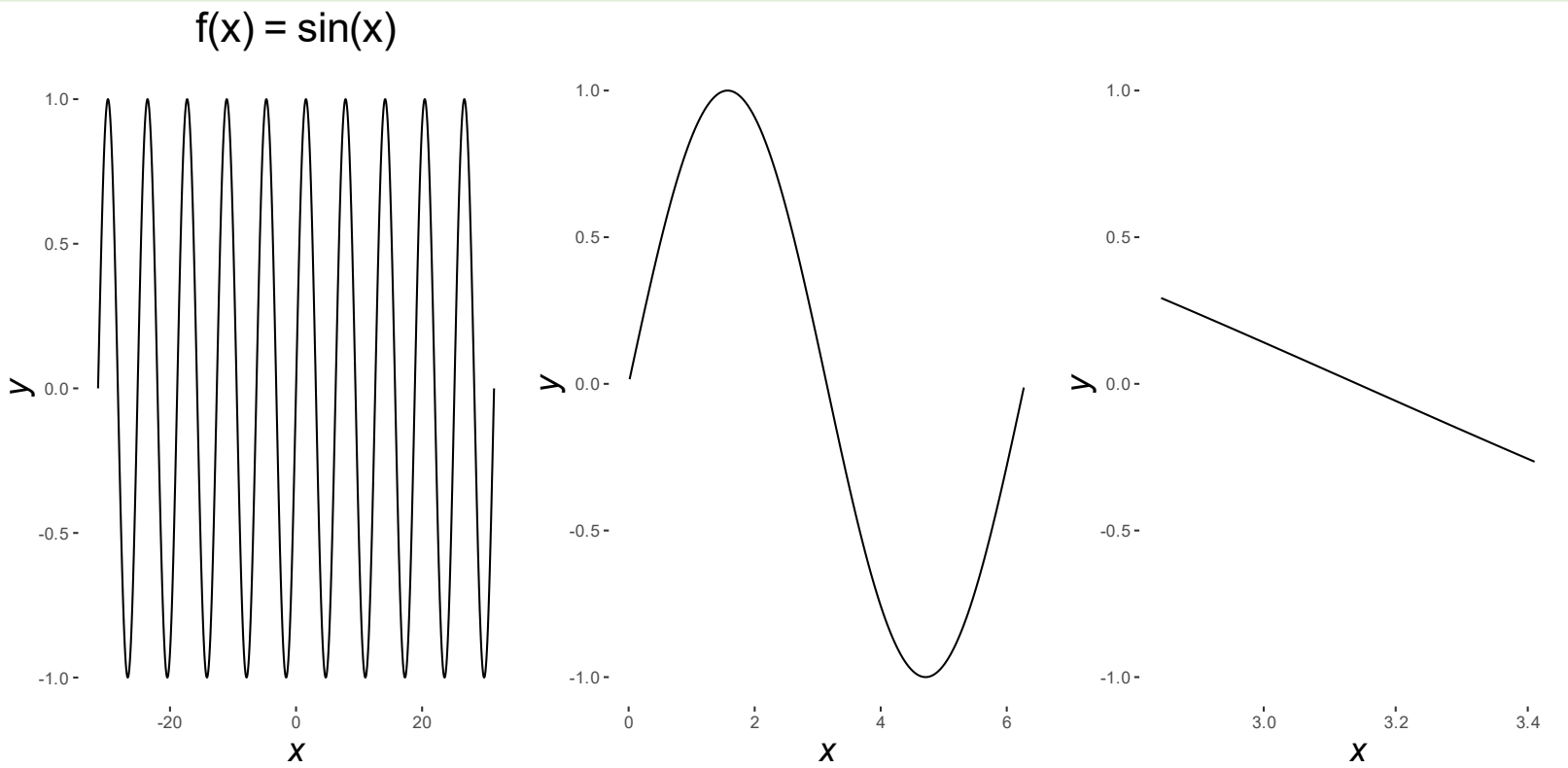
Local linearity

Could you model this curve with a linear function?



Local linearity

Could you model this curve with a linear function?



Local linearity

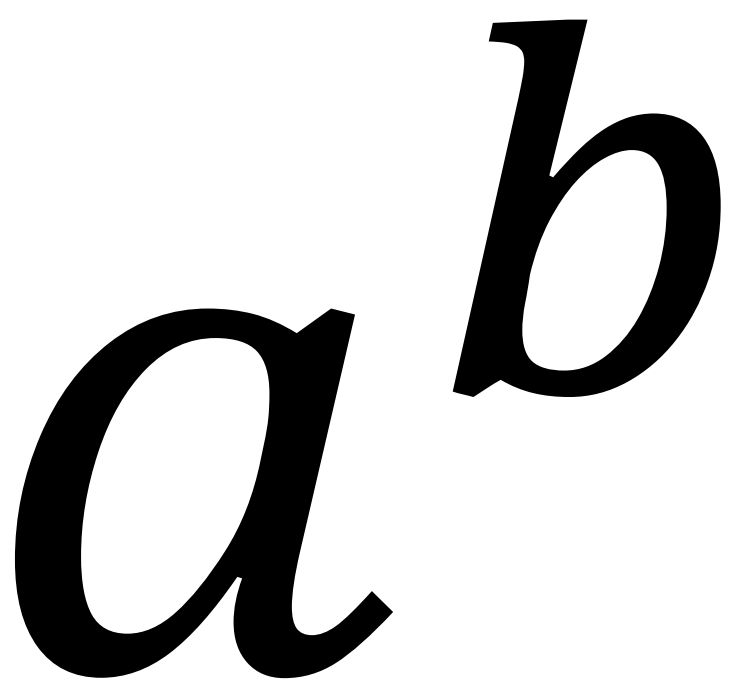
We are often justified in using a linear model, even when we know a relationship isn't linear:

- 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

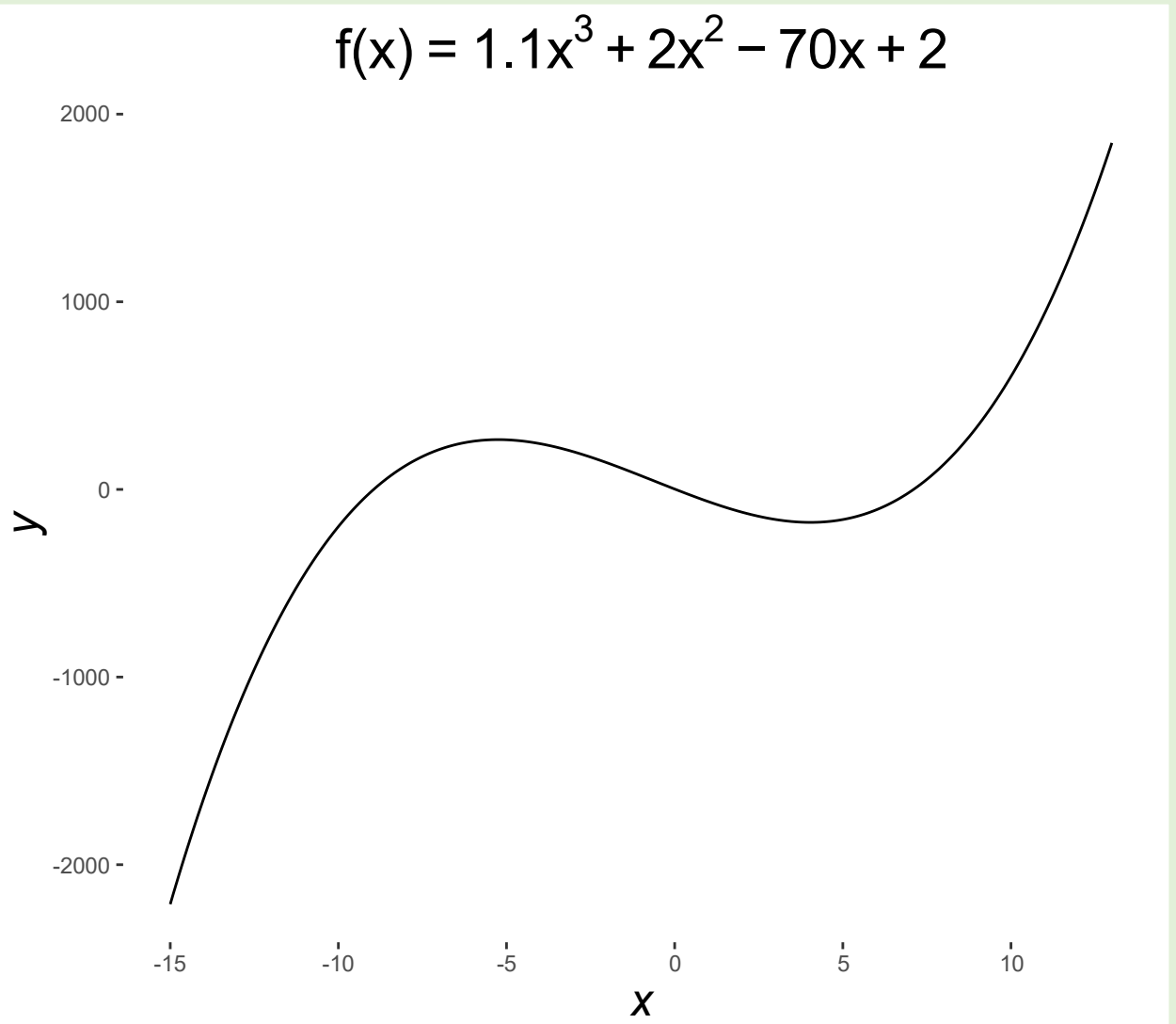
What are bases and exponents?	Notation convention:
Let a be a real number, and b be an integer: a is the base b is the exponent Exponentiation in this world means: "Multiply a by itself b times."	

Polynomial functions

Polynomial functions have non-negative integer powers:

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

Linear functions are a subset of polynomial functions.



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 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*!

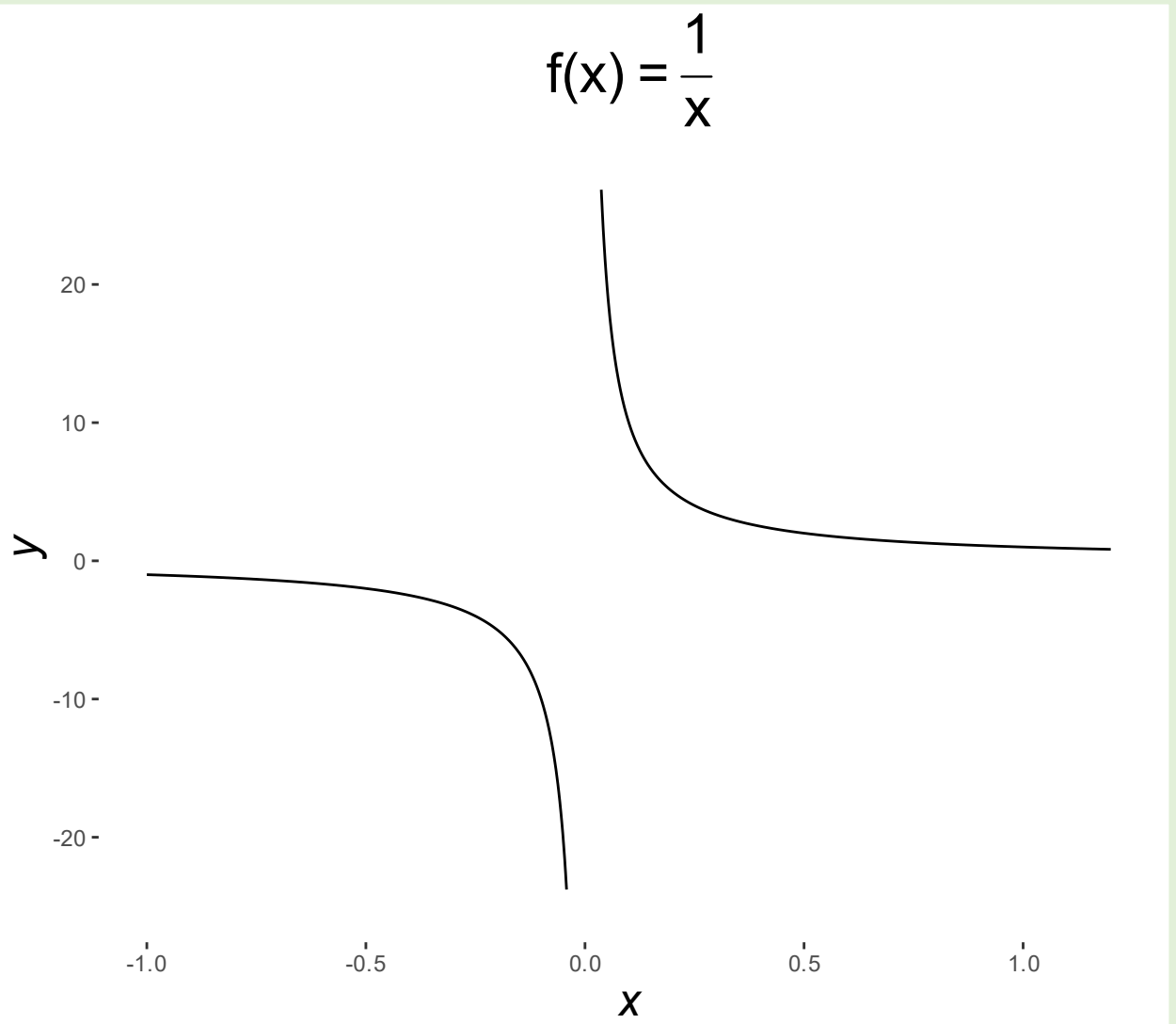
What does 'linear in the parameters' mean?

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

Rational functions

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

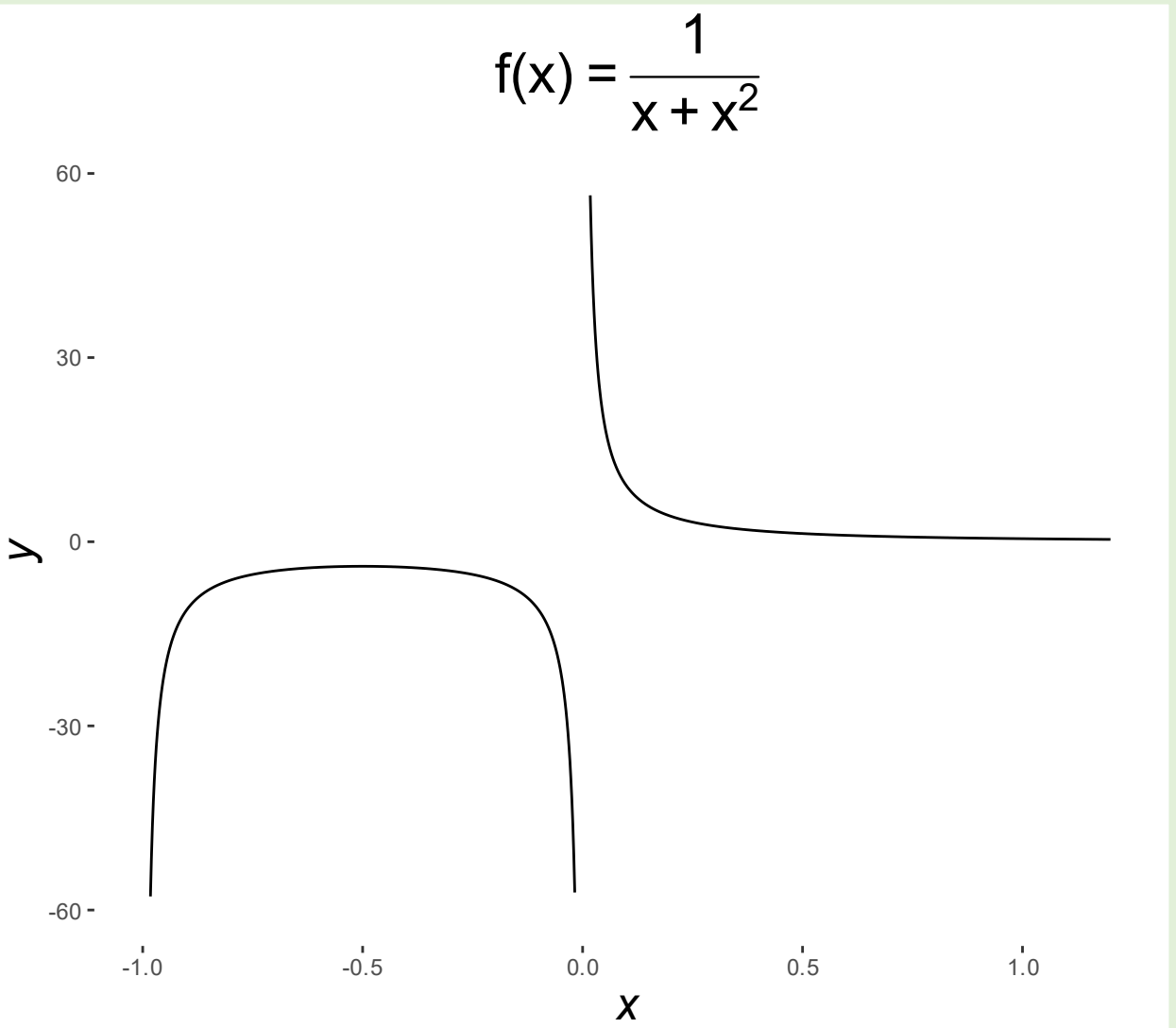
- Polynomial functions are a subset of rational functions.
- The square root function is *not* a rational function.
- Rational functions can be *discontinuous*: division by zero.



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



Fractional (rational) 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.	• I won't use this terminology, but you may find it in readings. • There are so-called power-law distributions (like the Pareto), but we won't be talking much abot these

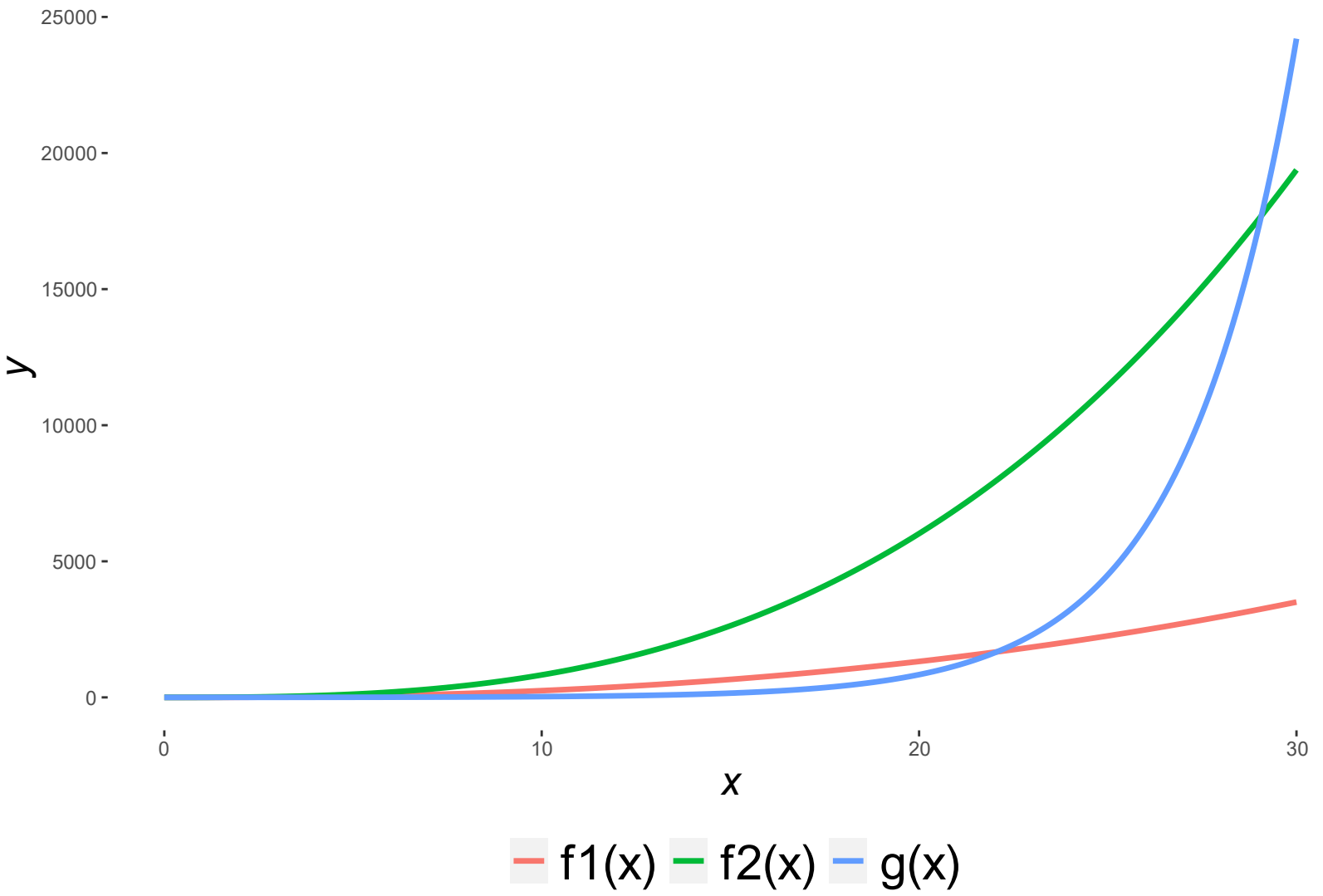
Power vs. exponential functions

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. exponetial 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.

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

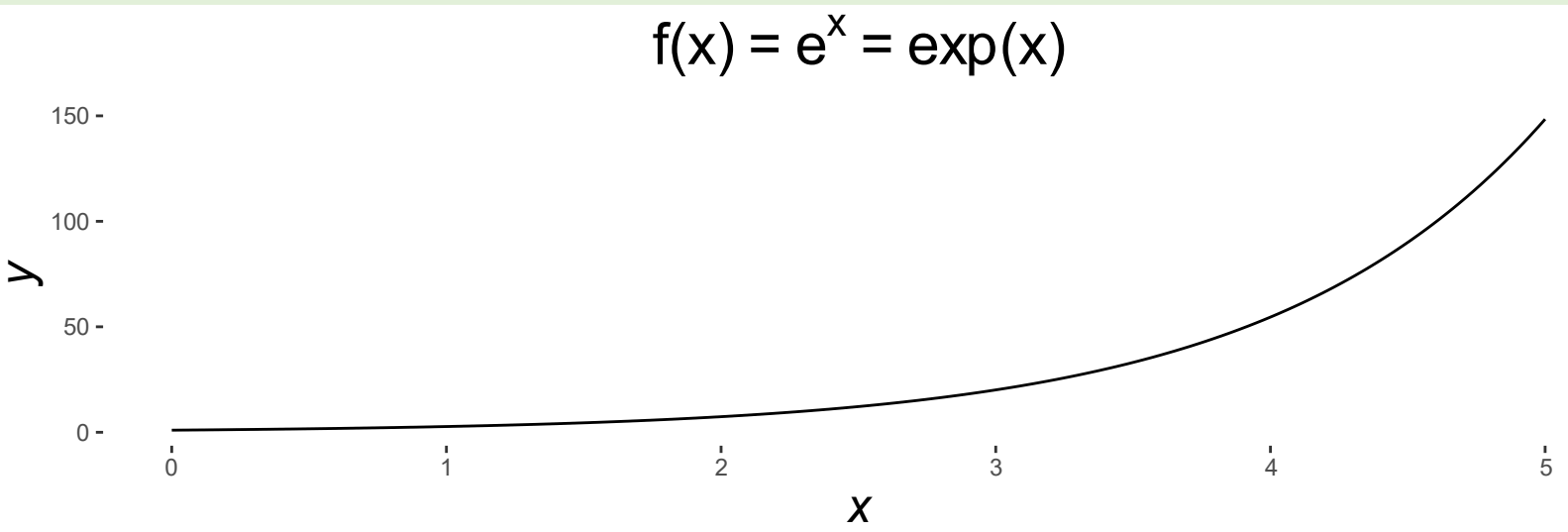
Expontials: the variable is the exponent; the base is a constant.

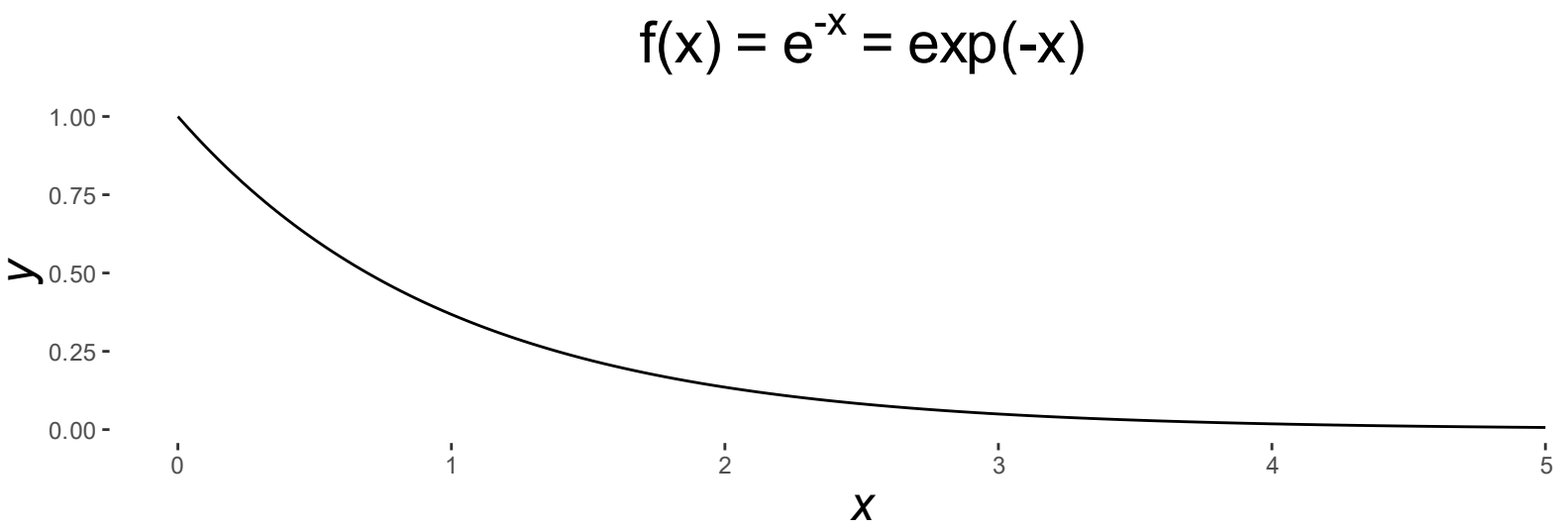
$$g(x) = 2.4^{x} - 0.5^{x}$$

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: $x^0 = 1$

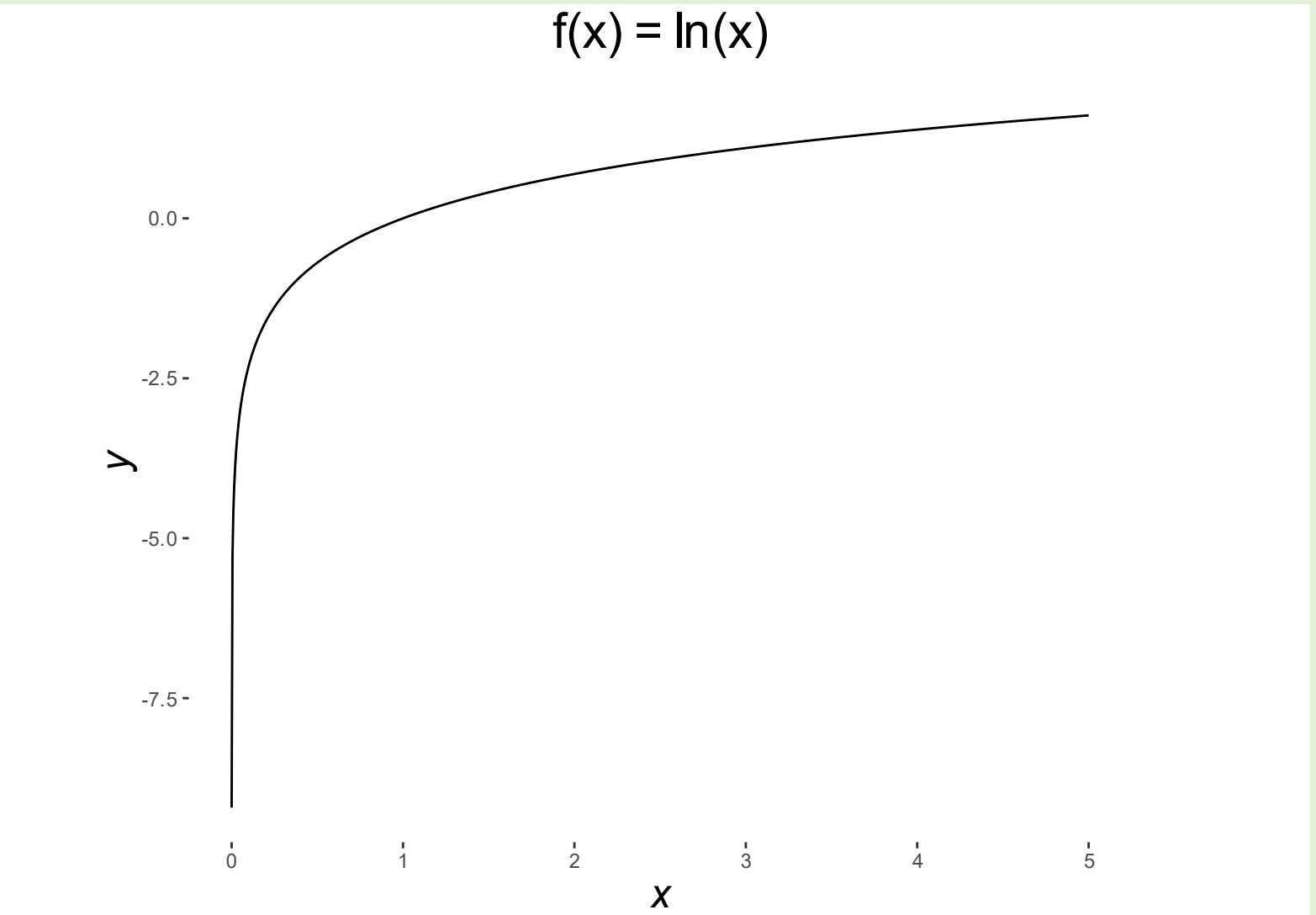




Logarithmic functions

Logarithmic functions are the inverse of exponential functions.

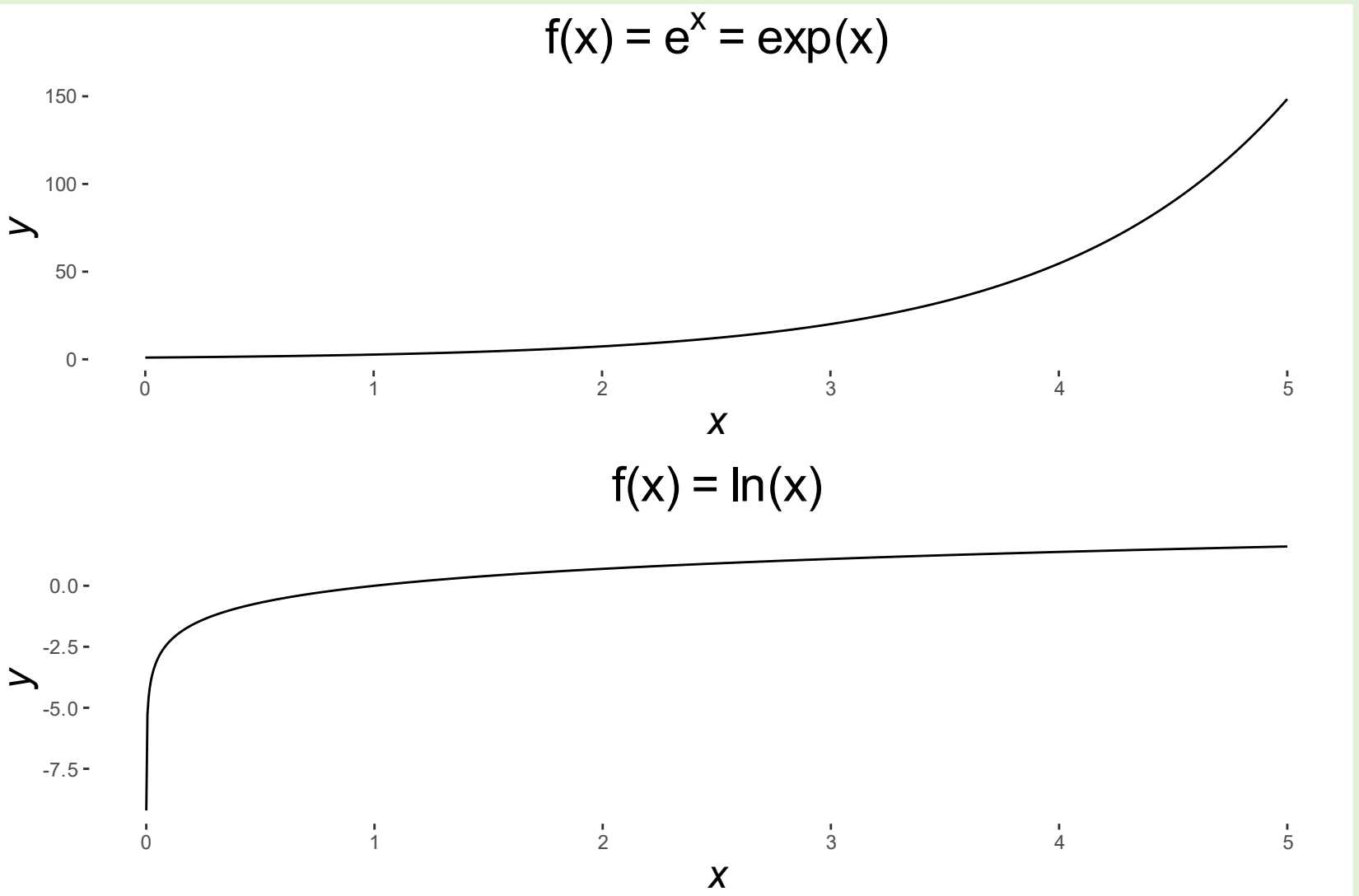
- Applying a logarithmic function *undoes* an exponential function.
- Logarithmic functions are slow-growing, but *not asymptotic*.



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

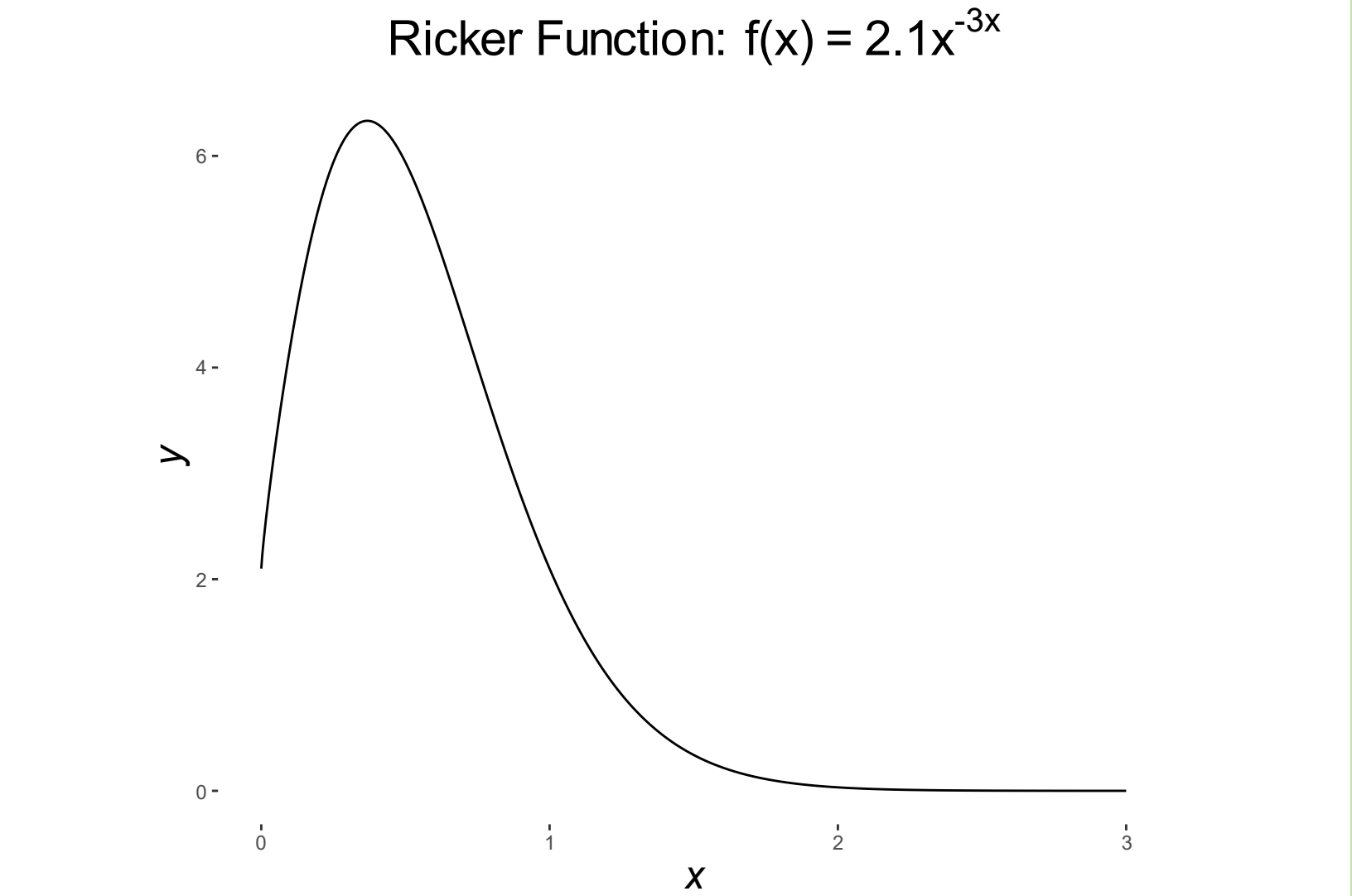


Hybrid functions

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

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

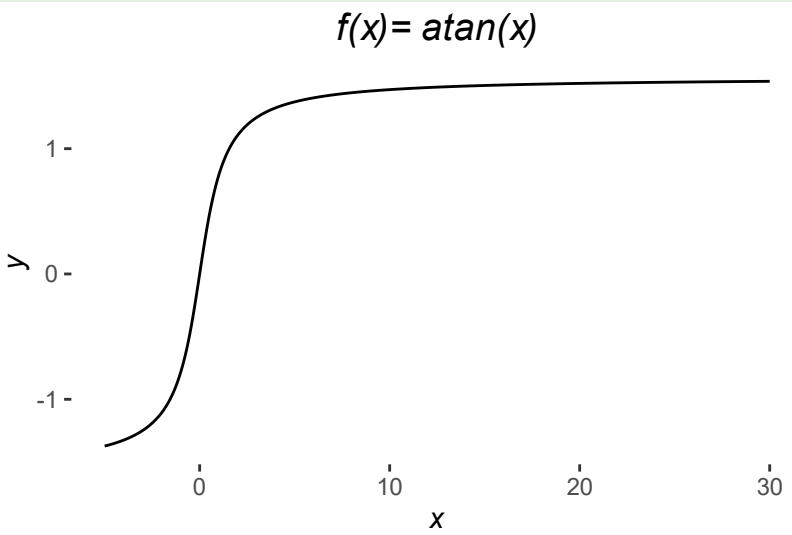
The Ricker function

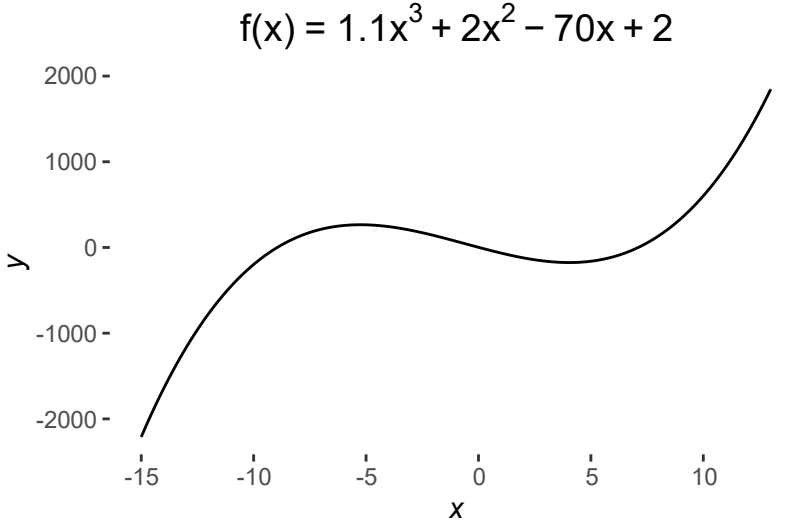


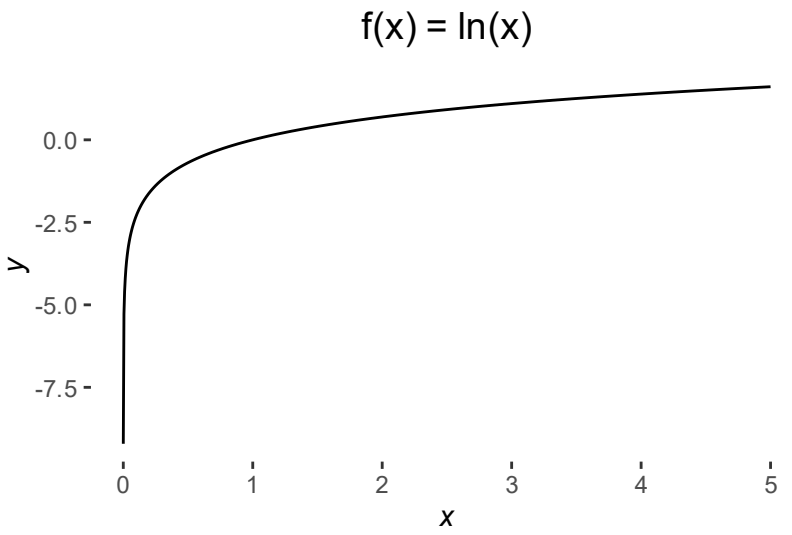
Graphical intuition

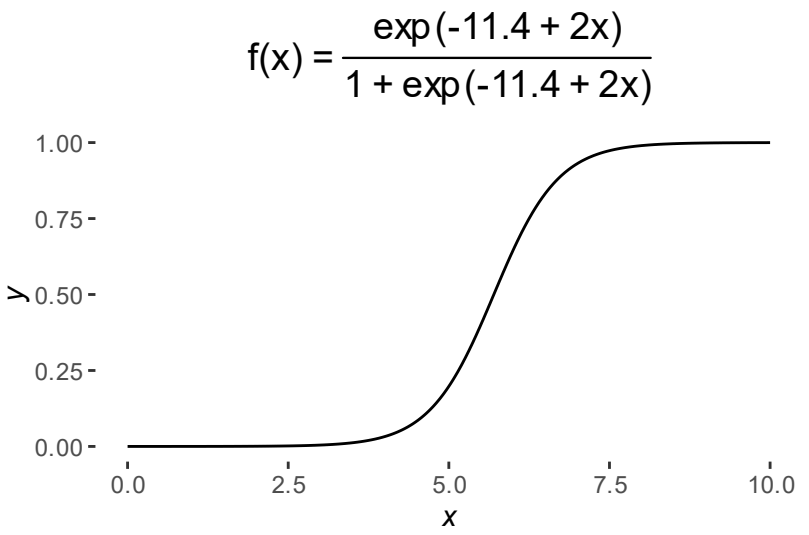
Function Terminology

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





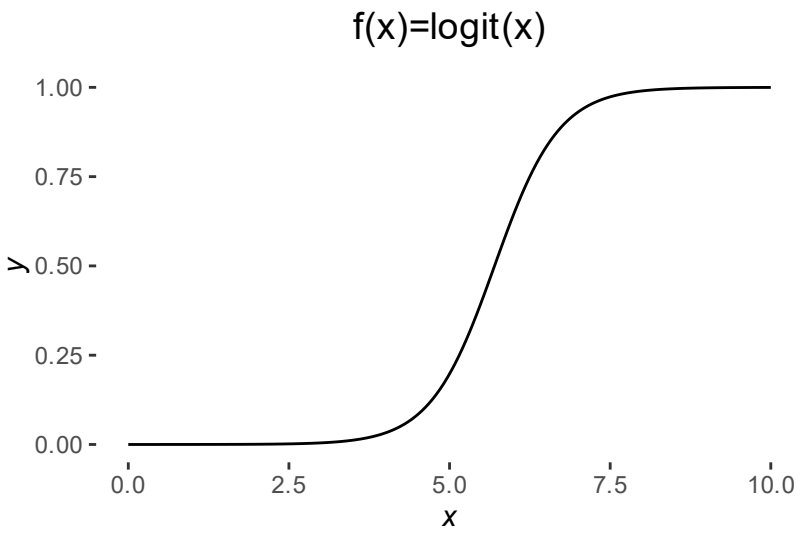


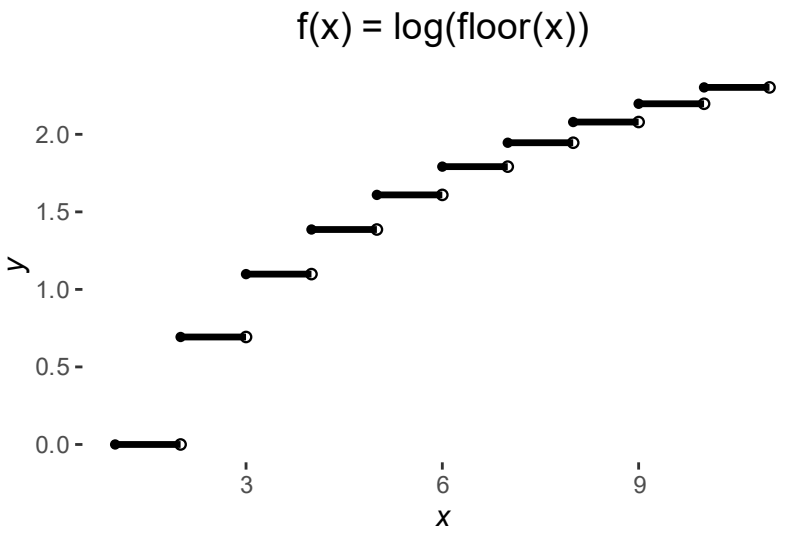


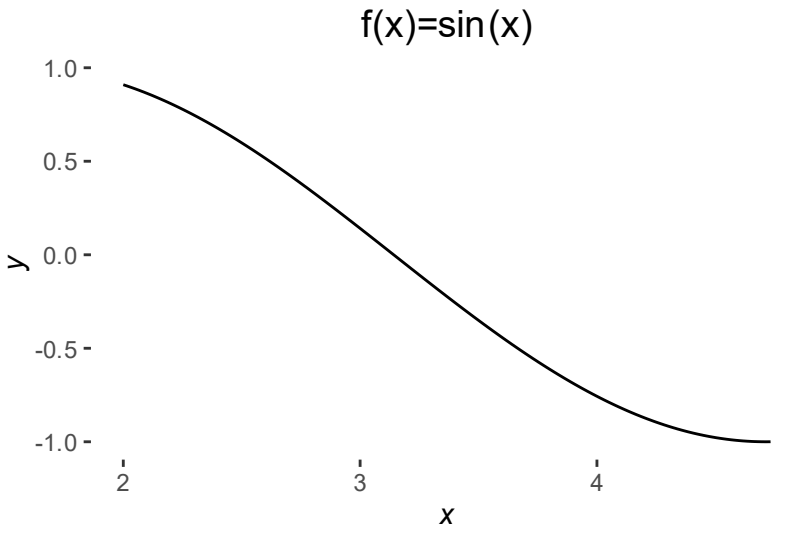
Graphical intuition

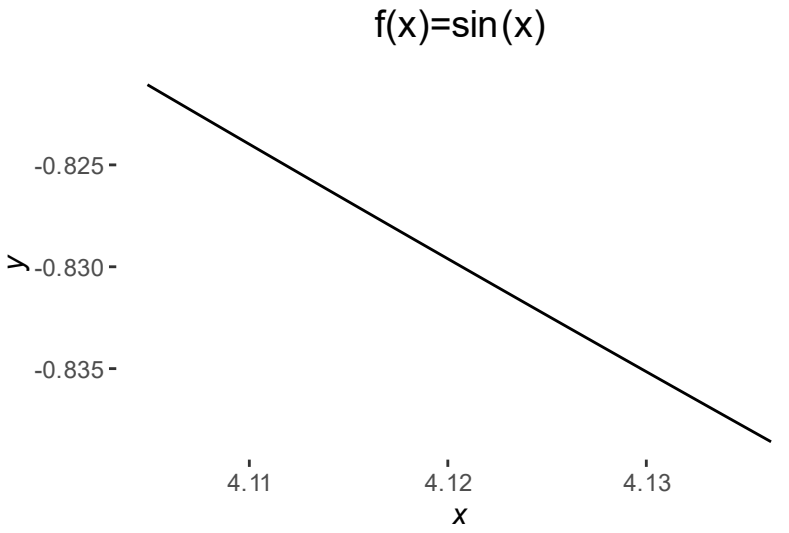
Function Terminology

- 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.









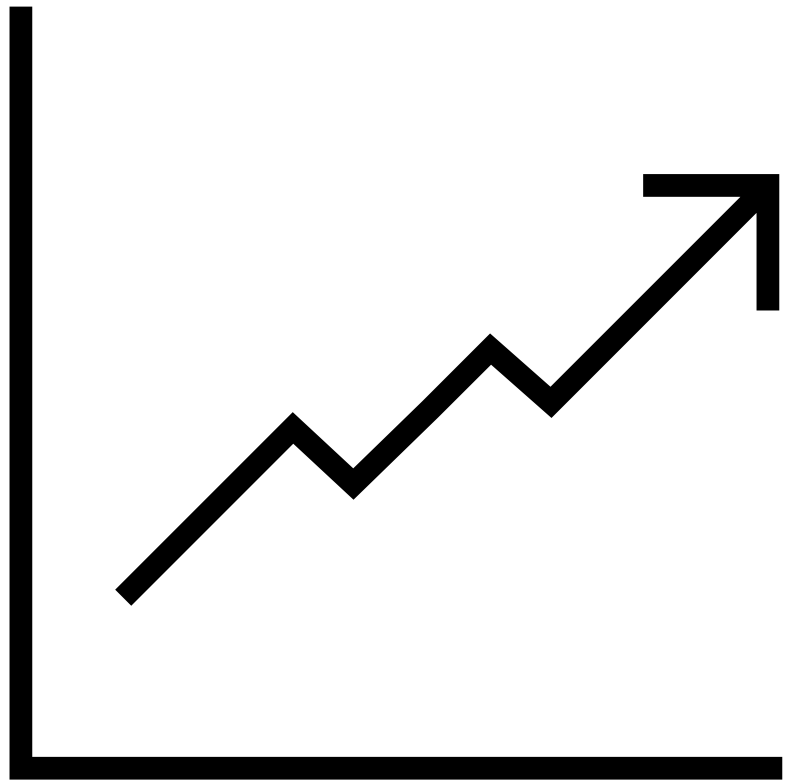
Associations

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

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

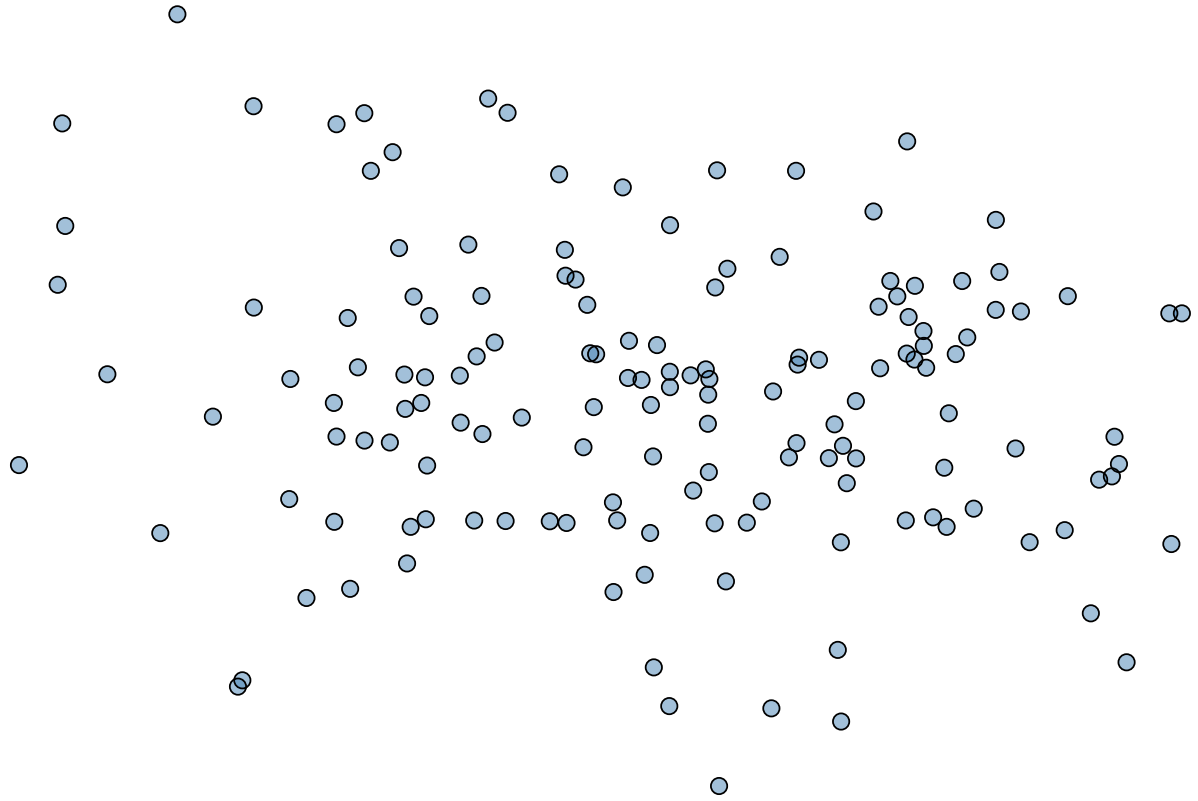
How can we *describe* an association?

- Qualitative and quantitative
- Numerically and graphically

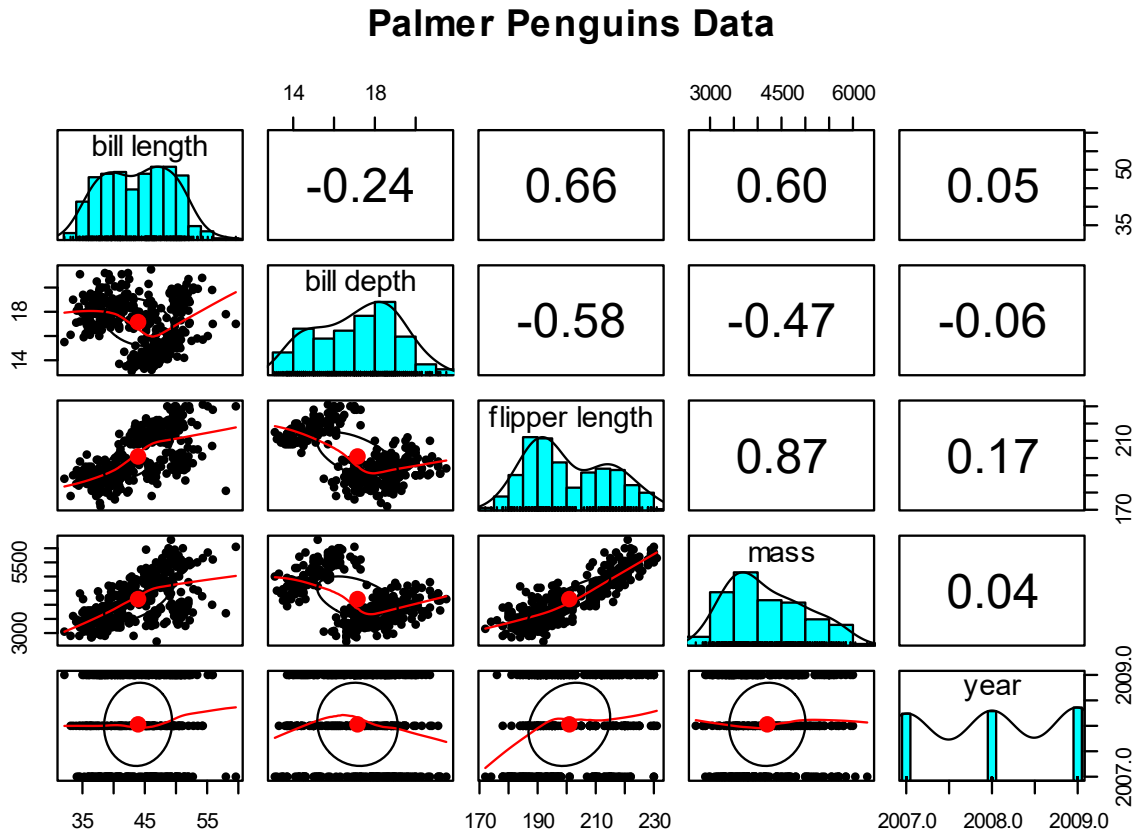


Associations: graphical exploration

Scatterplots are useful



Pairplots are even better!



Correlation Coefficients

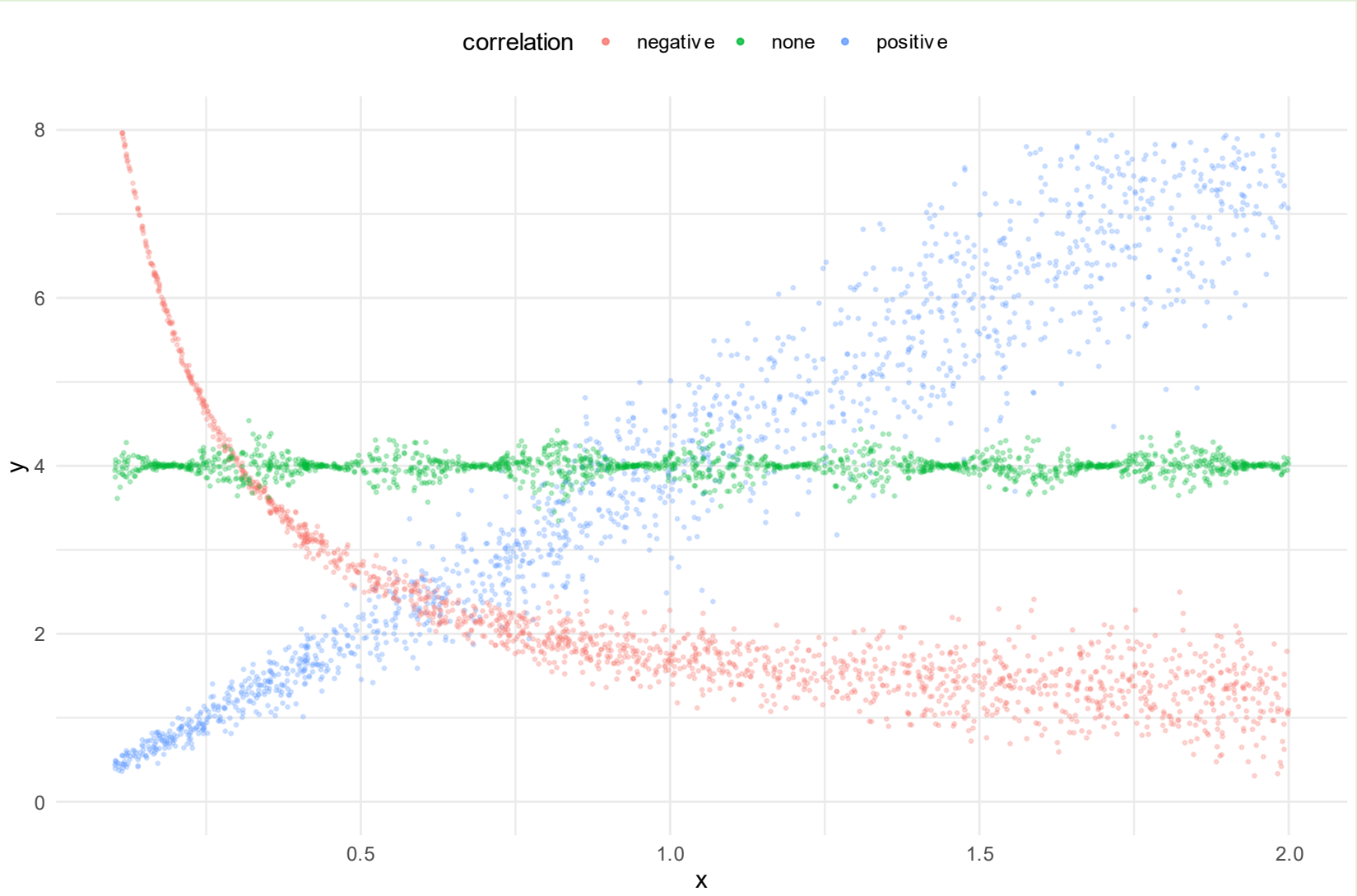
Correlations describe the strength of association between two variables	**Some limitations and caveats**

• 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.
	• 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.

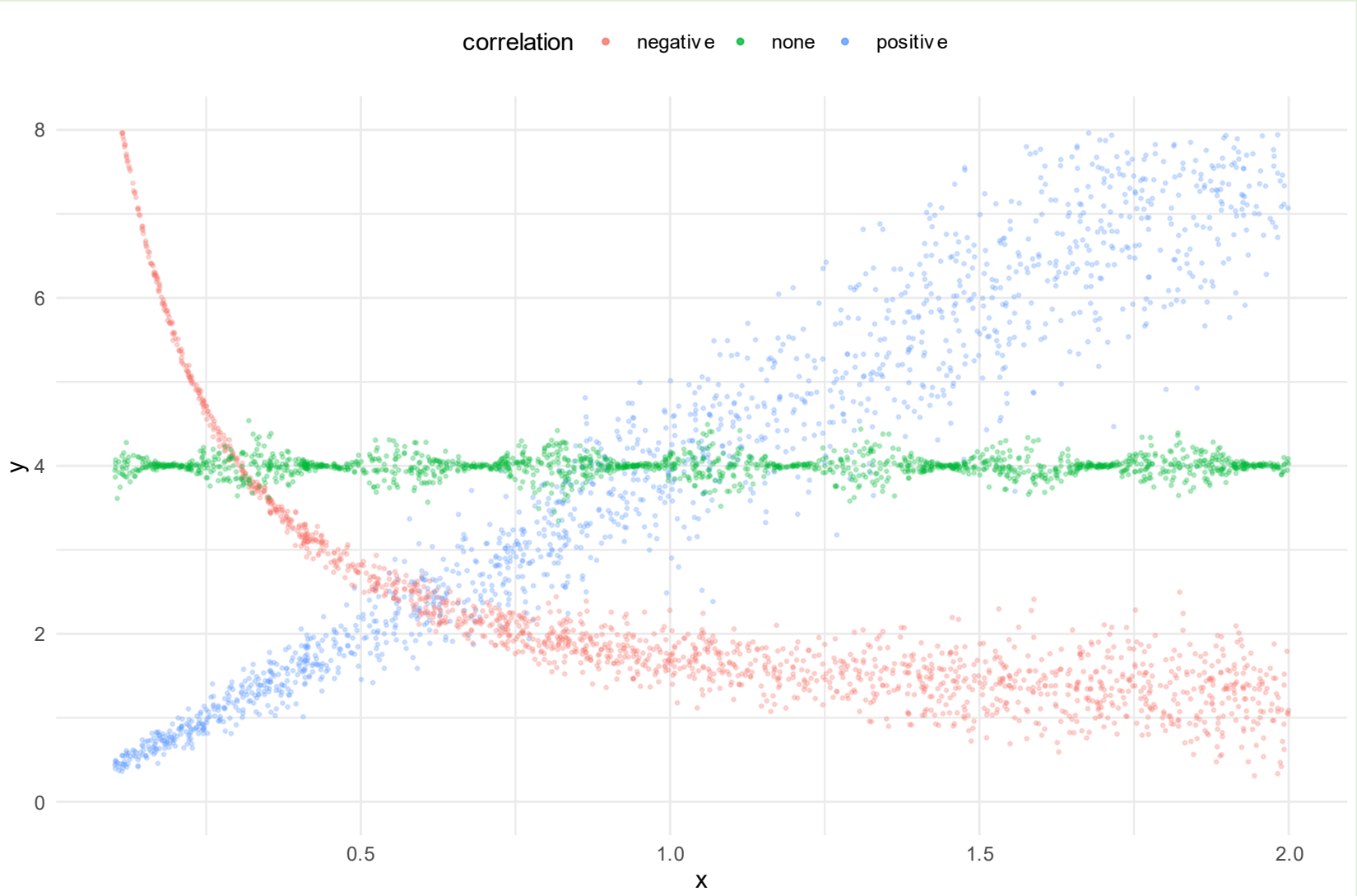
- Correlation ranges from -1 to 1:
- 1 indicates **perfect correlation**
  - Bivariate data lies exactly on a line of positive slope
- -1 indicates **perfect negative correlation**
  - Data lies exactly on a line with negative slope
- 0 Correlation: Points are **totally random** with respect to each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



Correlation – Information Perspective

Correlation Measures the Strength of the Association Between two Variables.

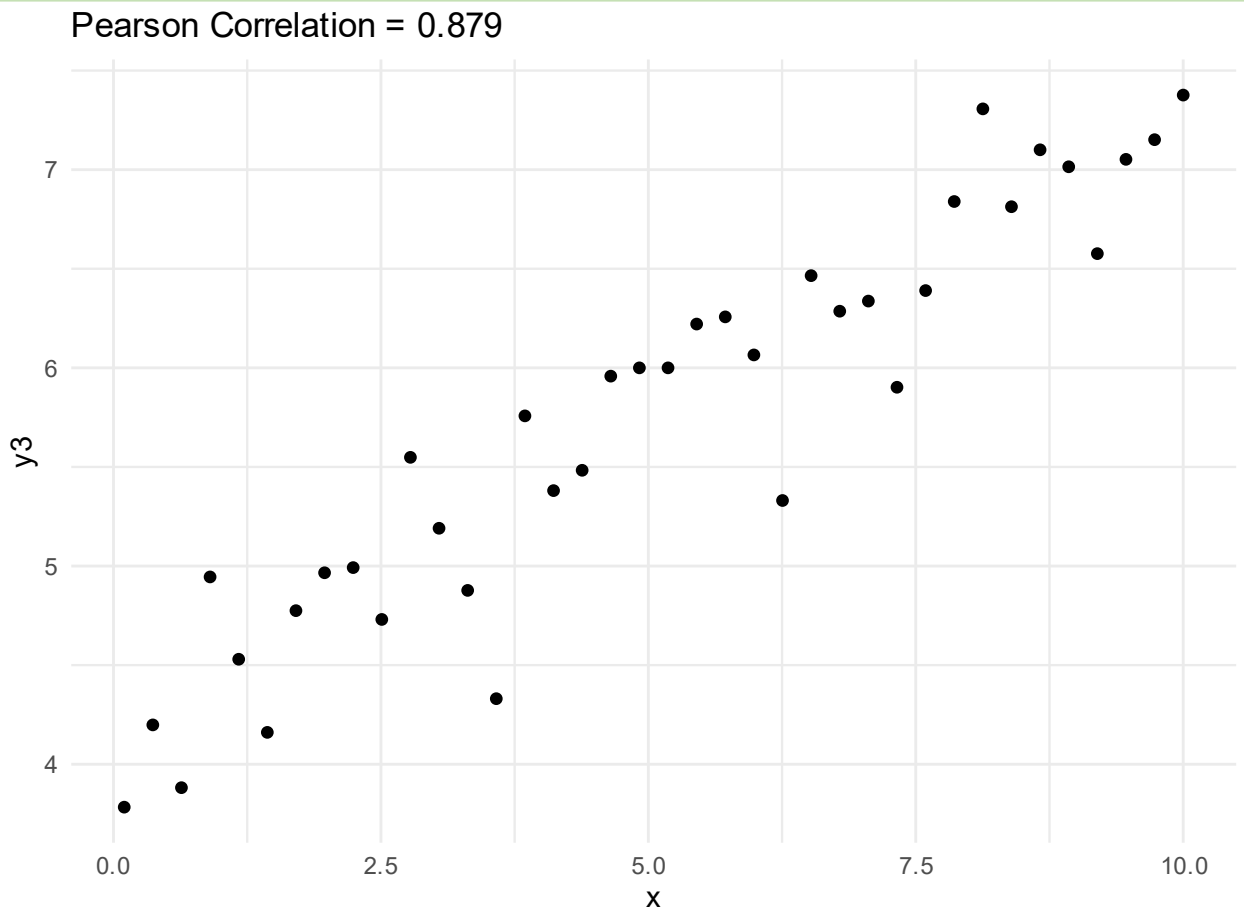
- Correlation ranges from -1 to 1:
- Corr 1: We can predict y from x perfectly. There's no noise, and as x increases, y increases. x tells us all we need to know about y.
- Corr -1: We can perfectly predict y from x, there is a negative relationship.with negative slope
- 0 Corr: X tells us nothing about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



Association: Numerical exploration

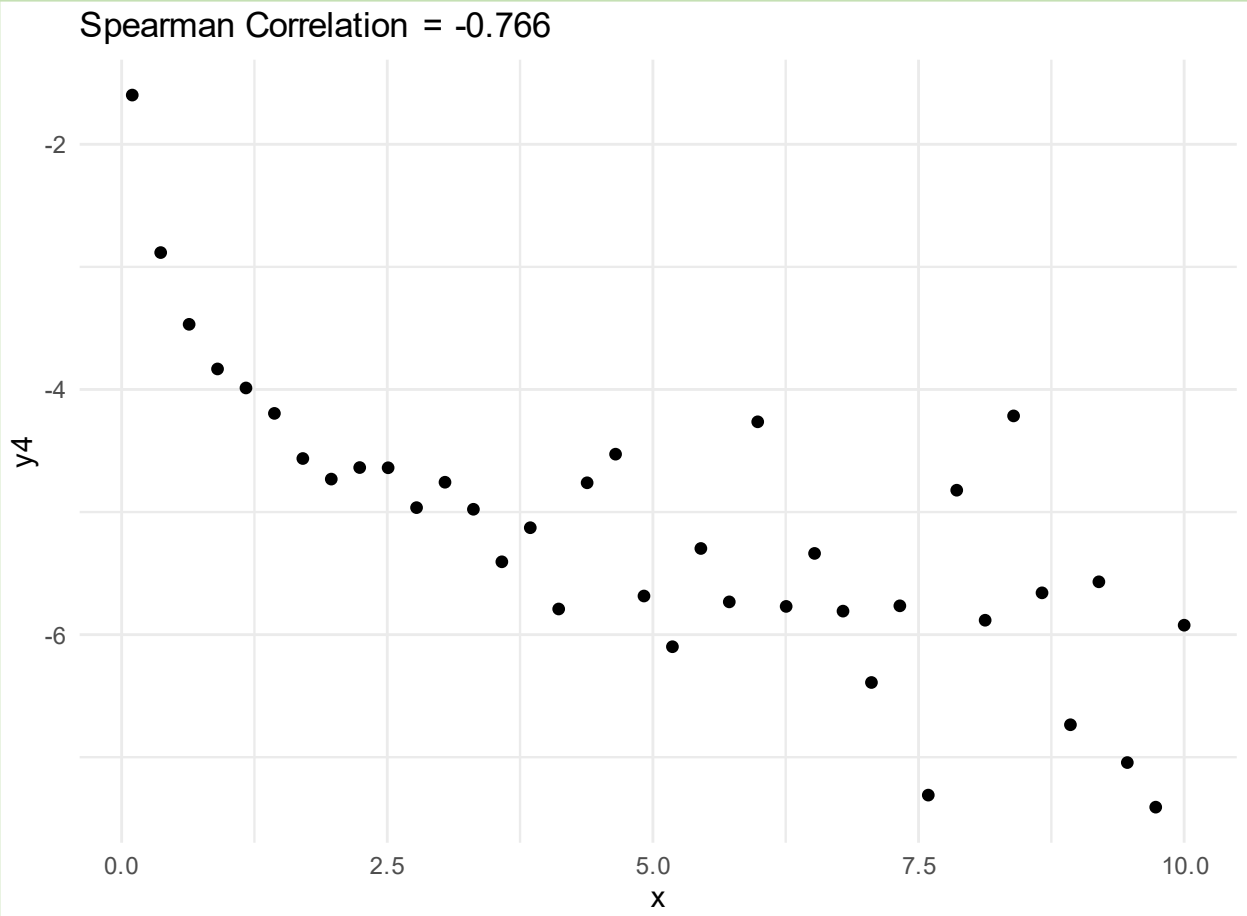
Quantifying linear association

Pearson Correlation



Quantifying monotonic association

Spearman Correlation

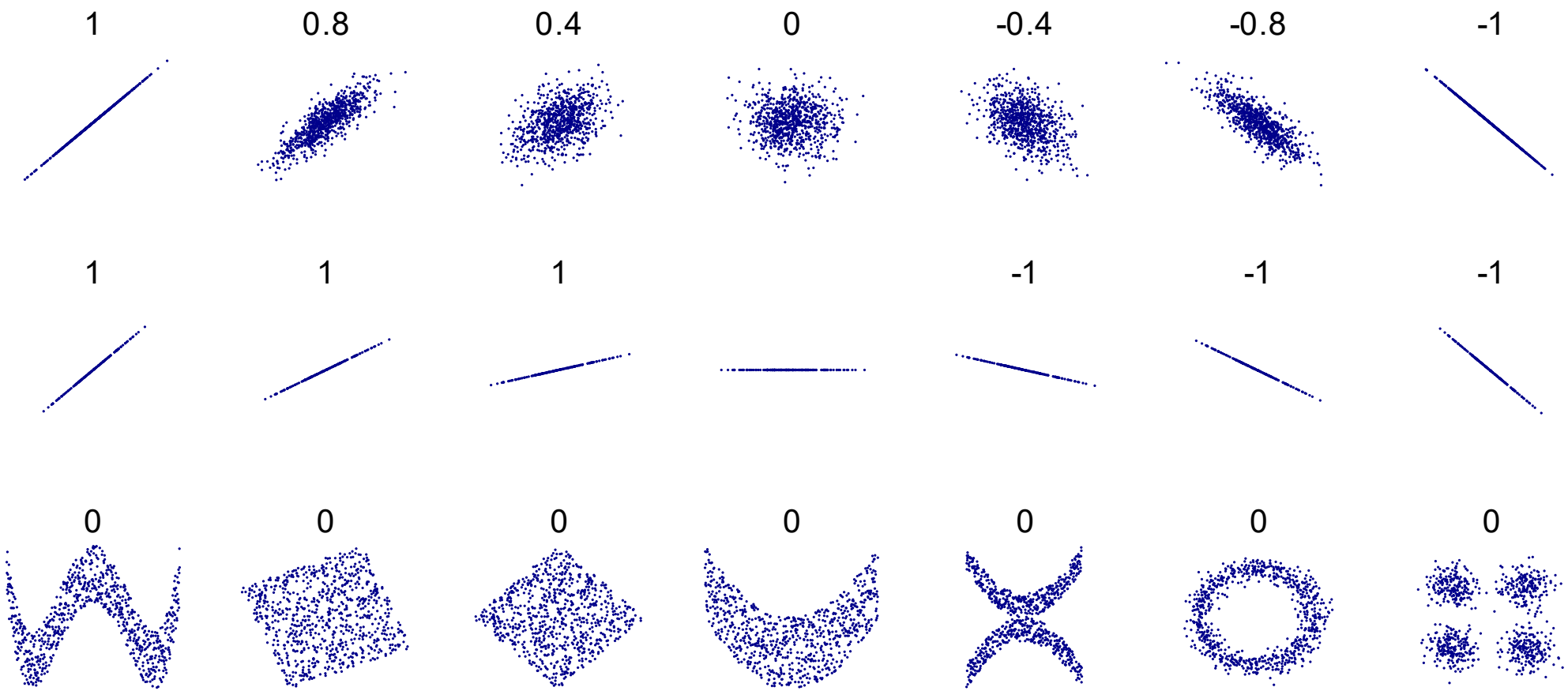


More complicated associations

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

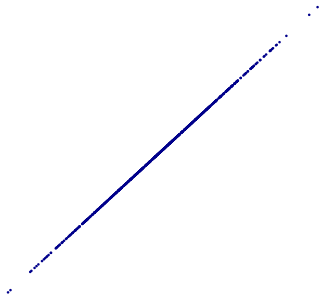
- Can you guess what the numbers represent?

Correlation Coefficients

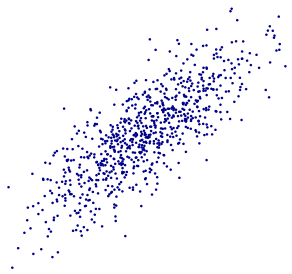


Denis Boigelot: public domain image

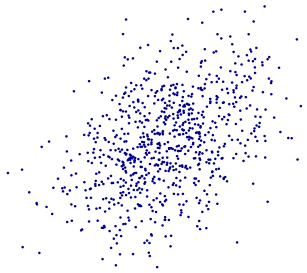
$s_{xy} = 2$, $r_{xy} = 1$



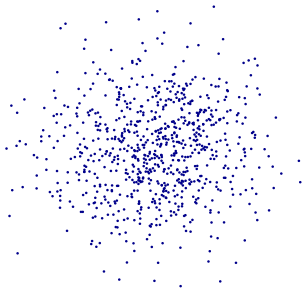
$s_{xy} = 1.6$, $r_{xy} = 0.8$



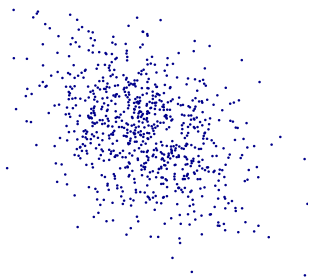
$s_{xy} = 0.9$, $r_{xy} = 0.4$



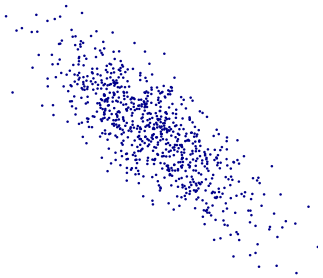
$s_{xy} = 0.1$, $r_{xy} = 0.1$



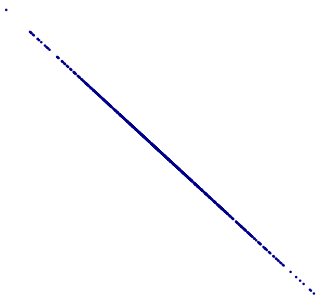
$s_{xy} = -0.8$, $r_{xy} = -0.4$



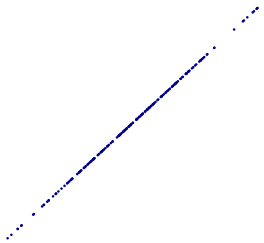
$s_{xy} = -1.6$, $r_{xy} = -0.8$



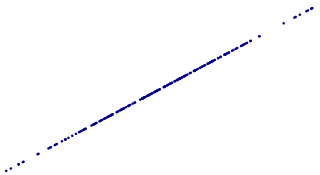
$s_{xy} = -2$, $r_{xy} = -1$



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



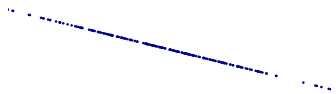
$s_{xy} = 1.8$, $r_{xy} = 1$



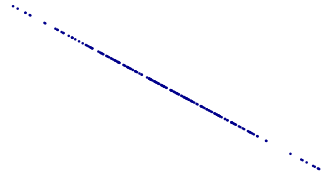
$s_{xy} = 1.1$, $r_{xy} = 1$

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

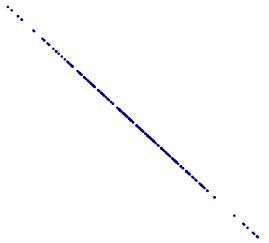
$s_{xy} = -1.1$, $r_{xy} = -1$



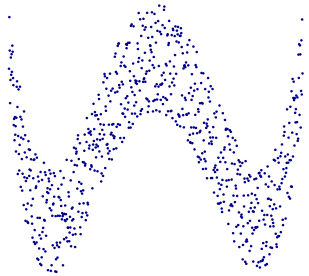
$s_{xy} = -1.8$, $r_{xy} = -1$



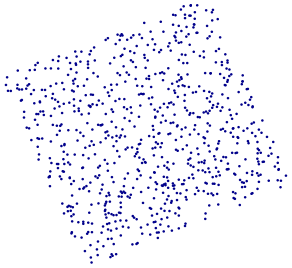
$s_{xy} = -2.1$, $r_{xy} = -1$



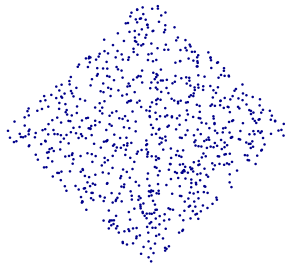
$s_{xy} = 0$, $r_{xy} = 0$



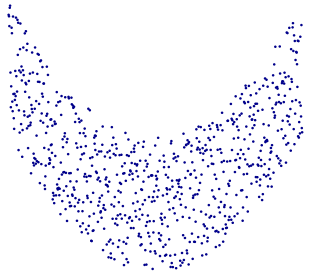
$s_{xy} = 0$, $r_{xy} = 0$



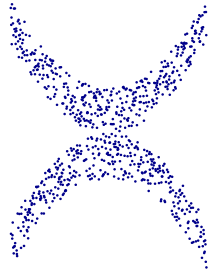
$s_{xy} = 0$, $r_{xy} = 0$



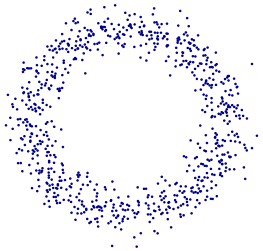
$s_{xy} = 0$, $r_{xy} = 0$



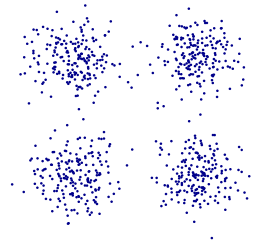
$s_{xy} = 0$, $r_{xy} = 0$



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



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



Data Dimensionality: How many variables do I have?

Data Dimensionality	Visualizing data
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'	• 1D: boxplots, histograms • 2D: conditional boxplots, scatterplots • 3D: coplots, 3D plots, 'slices' • 4D and is difficult or impossible • Multiple 3D 'panels' • 5D and is generally impossible

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

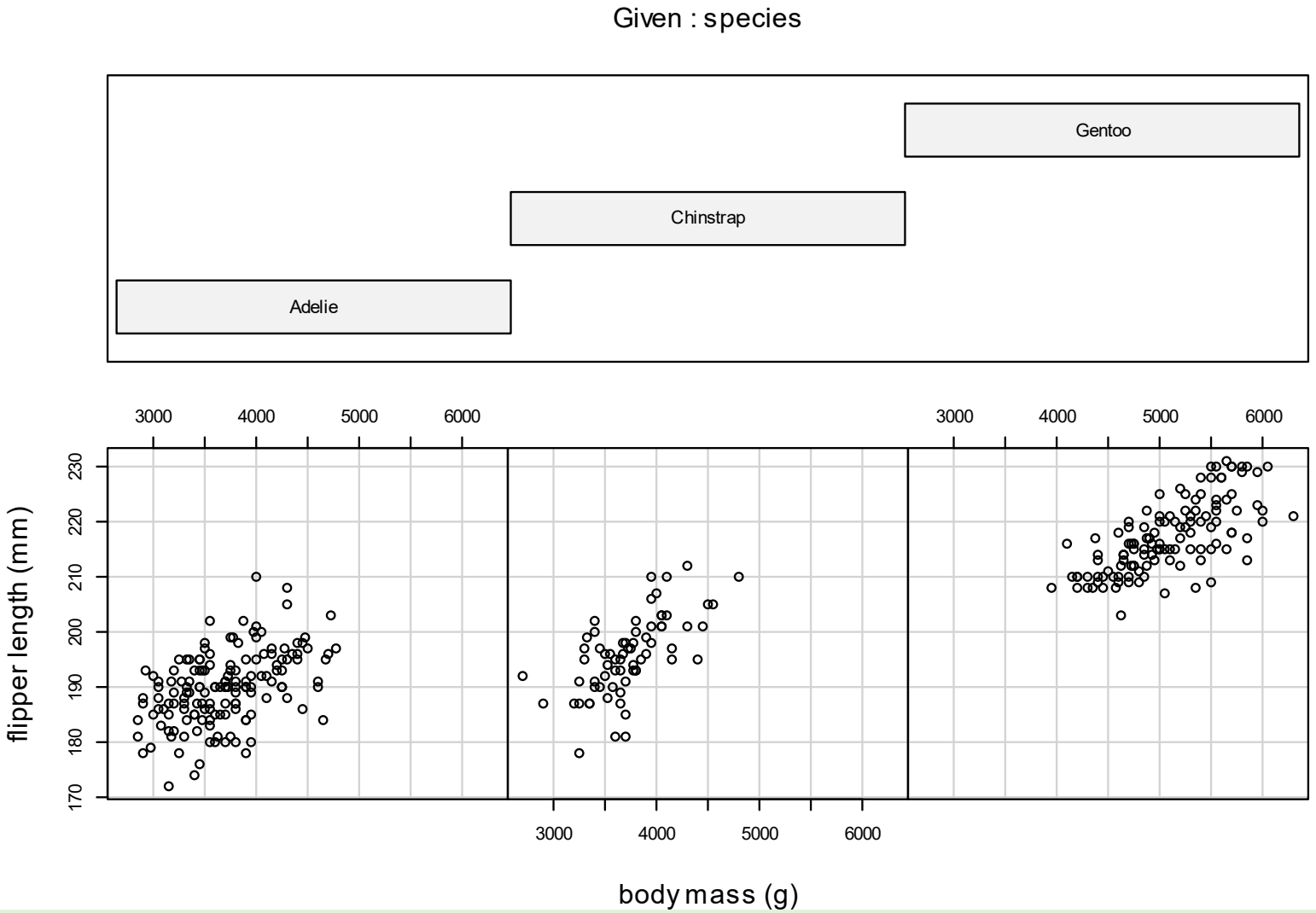
Visualizing 3D Data: Coplots

Coplot with a categorical variable

Each slice is a penguin species:

- **Adelie**
- **Chinstrap**
- **Gentoo**

What can you see?



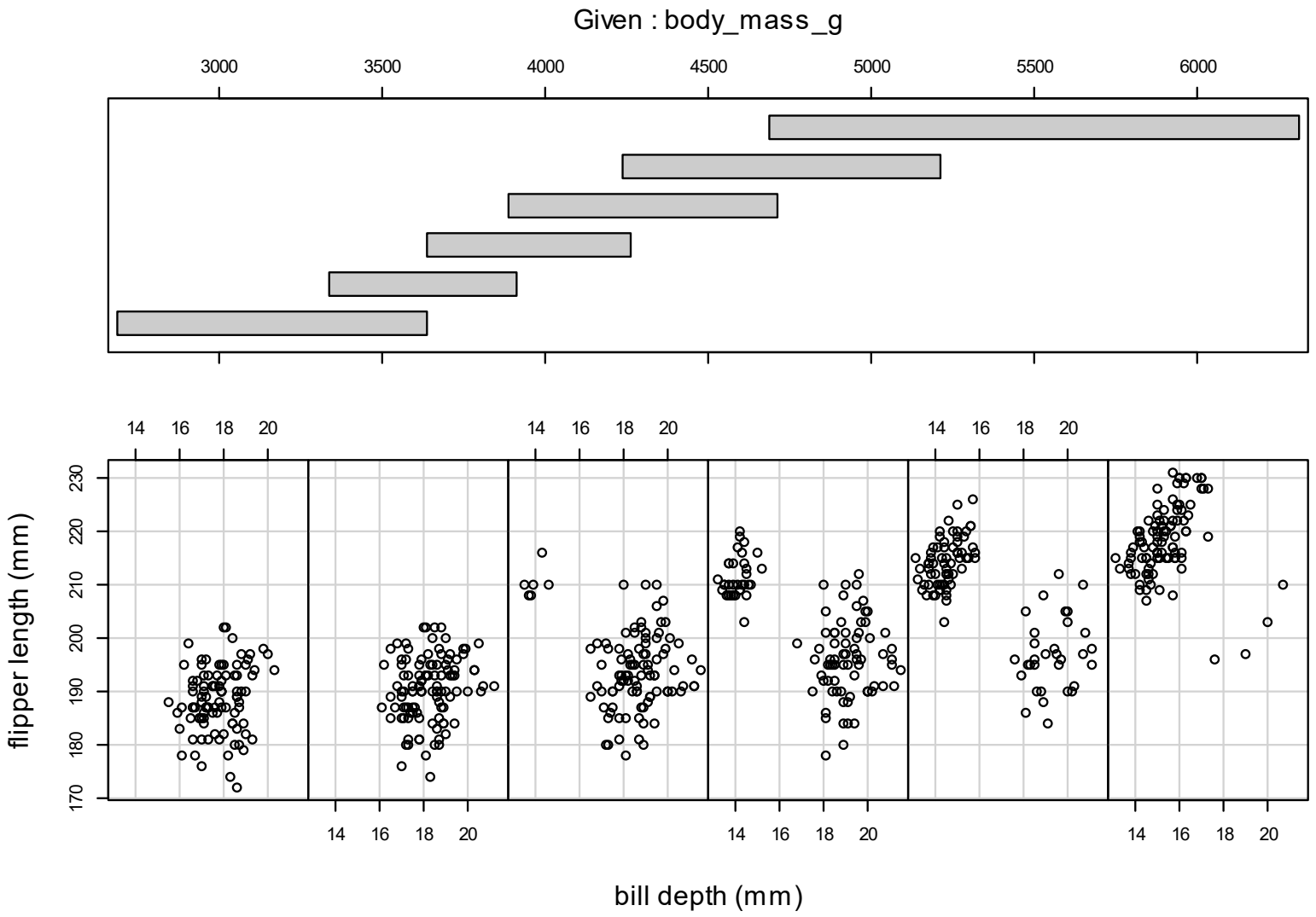
Visualizing 3D Data: Coplots

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?



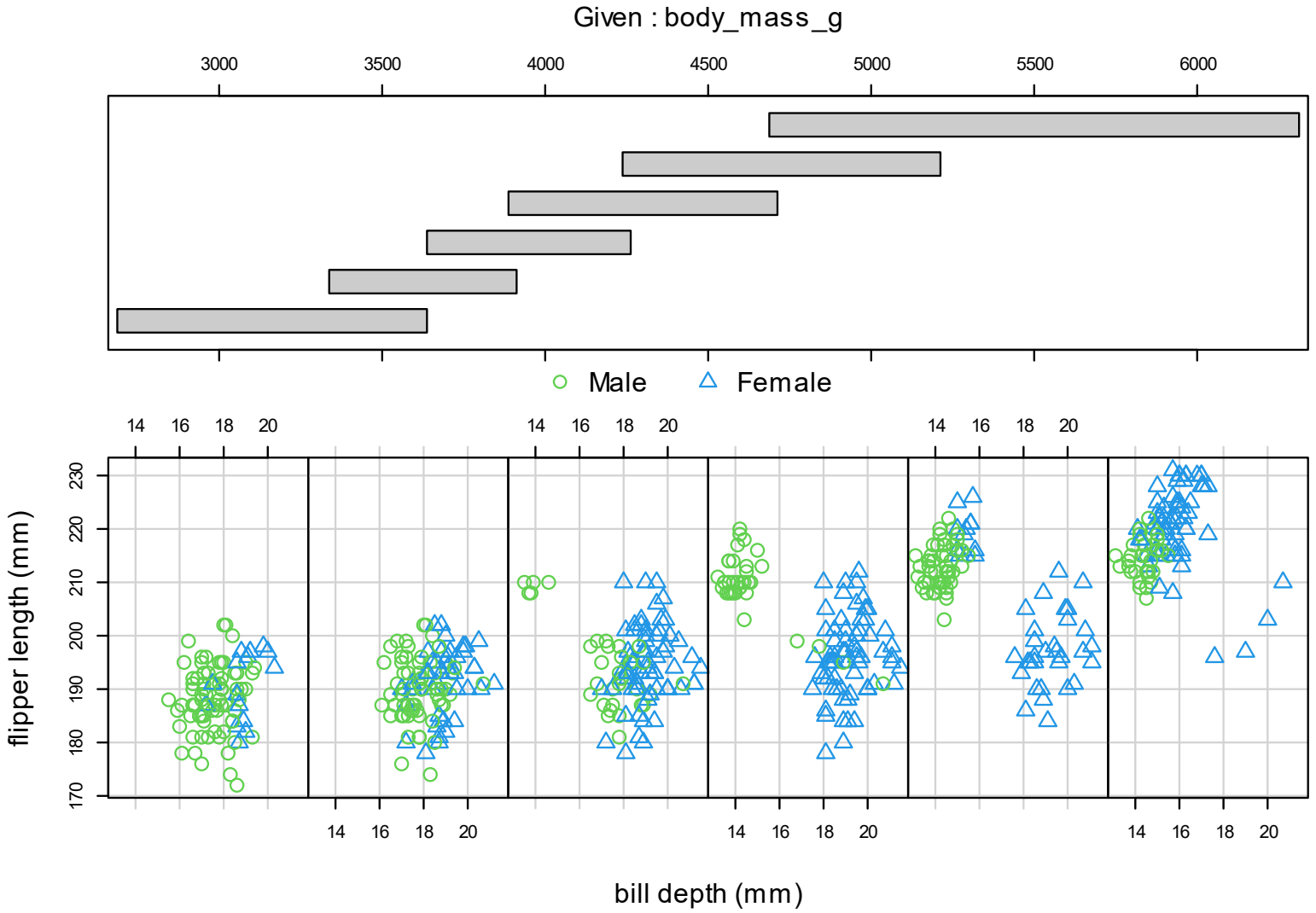
4D Slice Plots: Point Color and Shape

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?



4D example: Modeling Mountain Pine Beetle epidemics

4-dimensional data

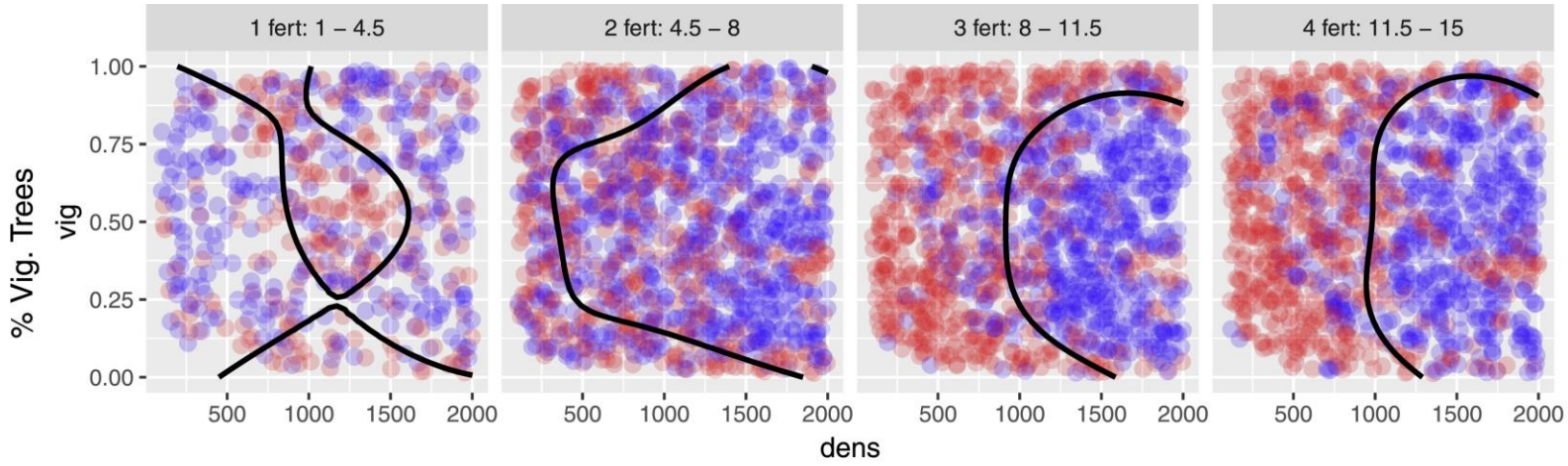
Three model parameters:

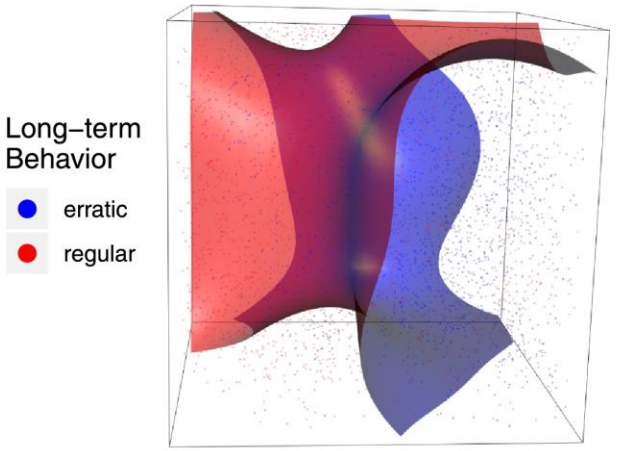
1. Beetle fertility
2. Tree vigor
3. Tree density

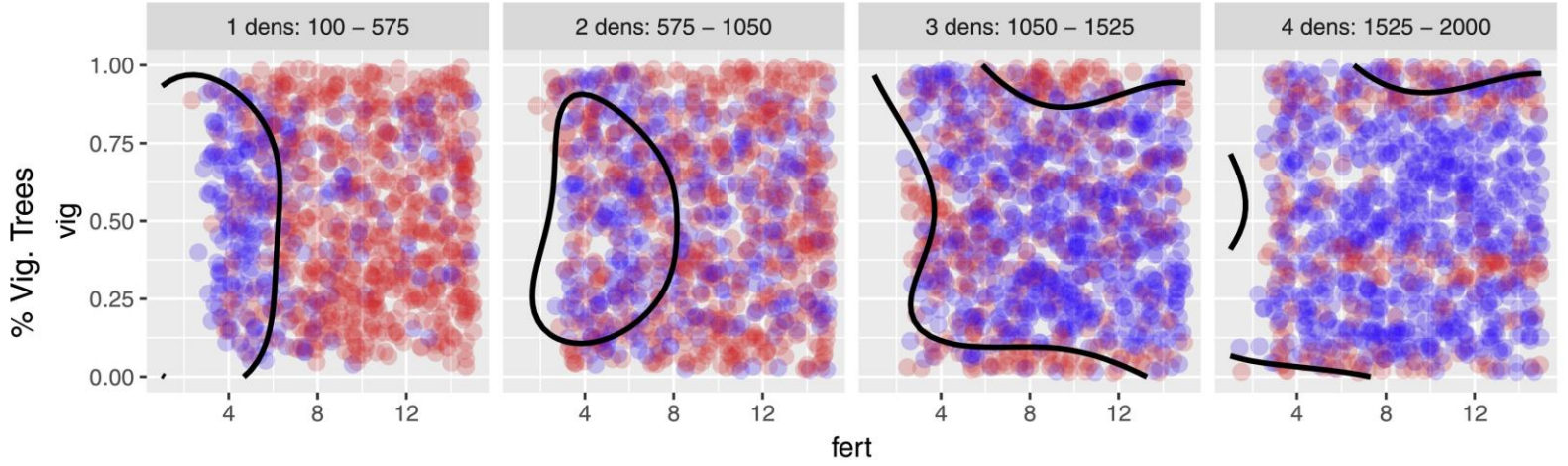
Two response types:

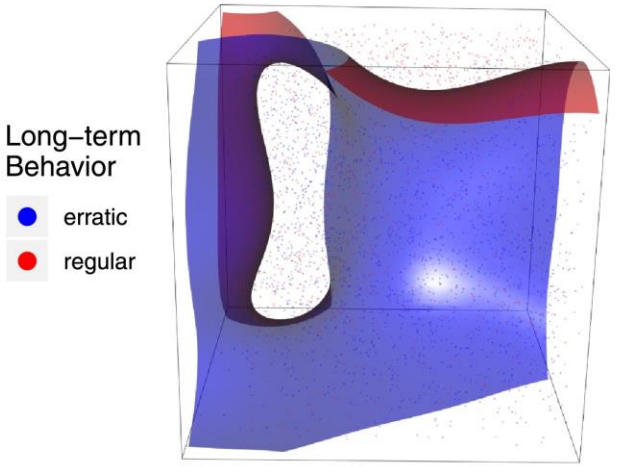
1. Long term epidemic behavior - categorical: erratic or regular epidemics
2. Epidemic proportion - continuous: long-term average percent of epidemic area

4D Slice Plots: Slices + Continuous Response as Color

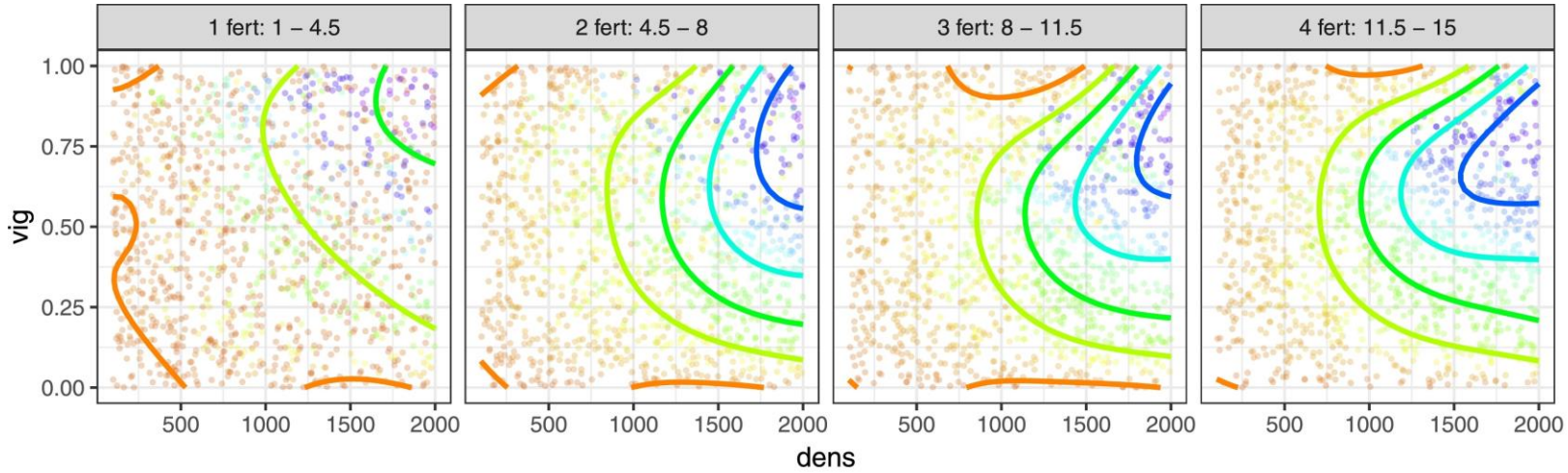


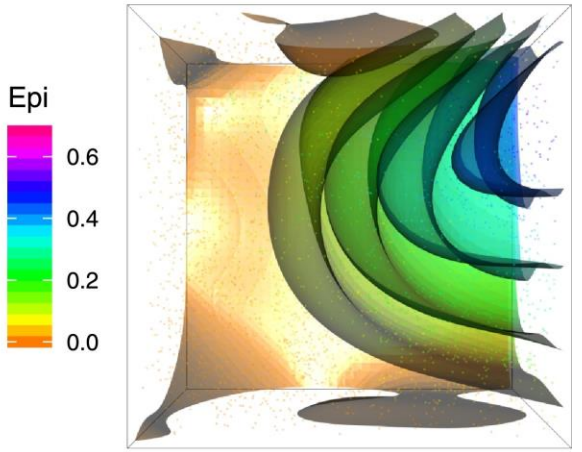


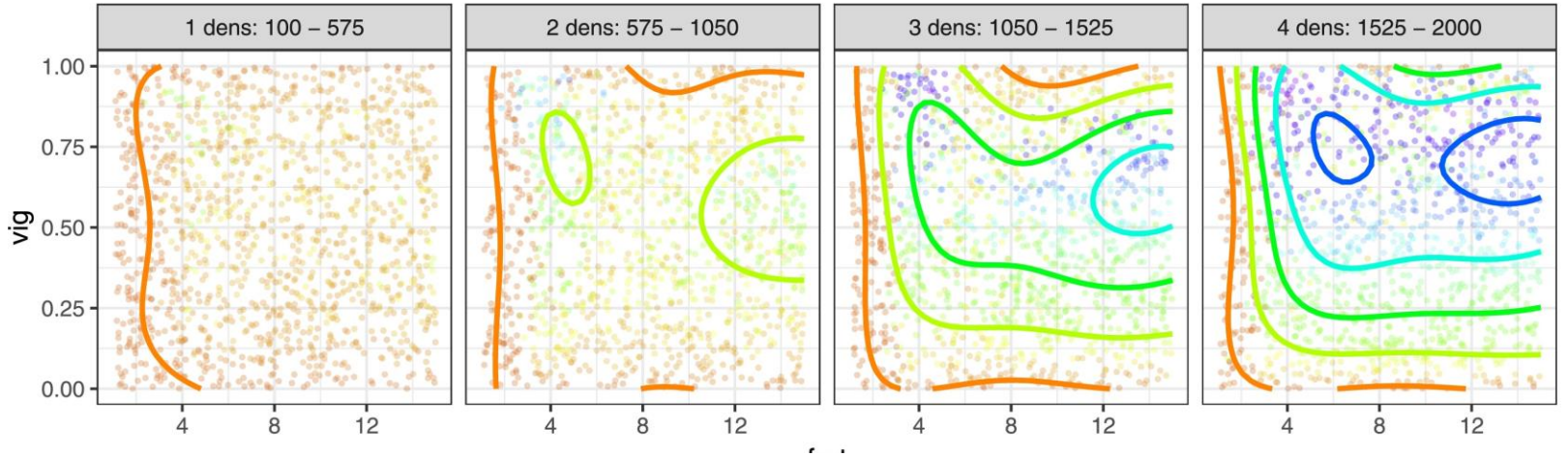


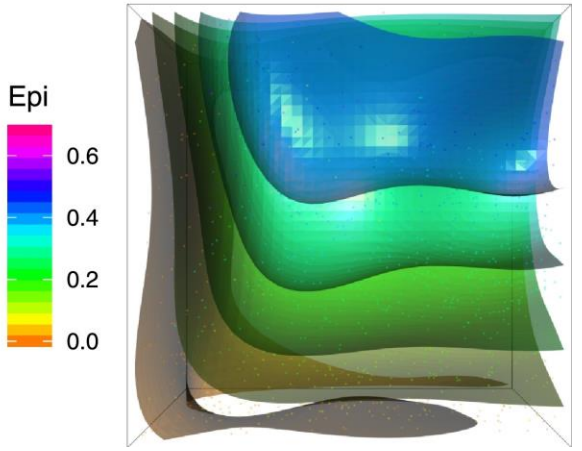


4D Slice Plots: Slices + Continuous Response as Color









Graphical Exploration Recap

- Associations
- Tools to describe associations
  - Spearman and Pearson correlation
  - Graphical exploration
- Data dimensionality and plotting

Zoom Poll

- Open up Zoom to the course channel – it's time for a poll!

Probability Distributions 1

General Concepts

Key probability terms and concepts

- Inference with the dual model paradigm
- What is a distribution?
- Event, domain, sample space
- 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

Stepping back: What do we need to do inference?

We need the Dual Model Paradigm to do inferential statistics.

- 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?

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?	**Probability Distributions**

• 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
	• 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

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 take a detour to talk about probability theory…

Parametric and Empirical Distributions

Parametric distributions are defined by mathematical *functions*	**Empirical distributions are computed from *observations*.**
• 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.	• There is no analytical function, but we can compare empirical distributions to parametric distributions. • Useful for comparing null and alternative hypotheses

Probability Distribution Functions

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.

Two other types of functions are used to describe distributions:

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

Density or Mass Function: PDFs & PMFs

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.

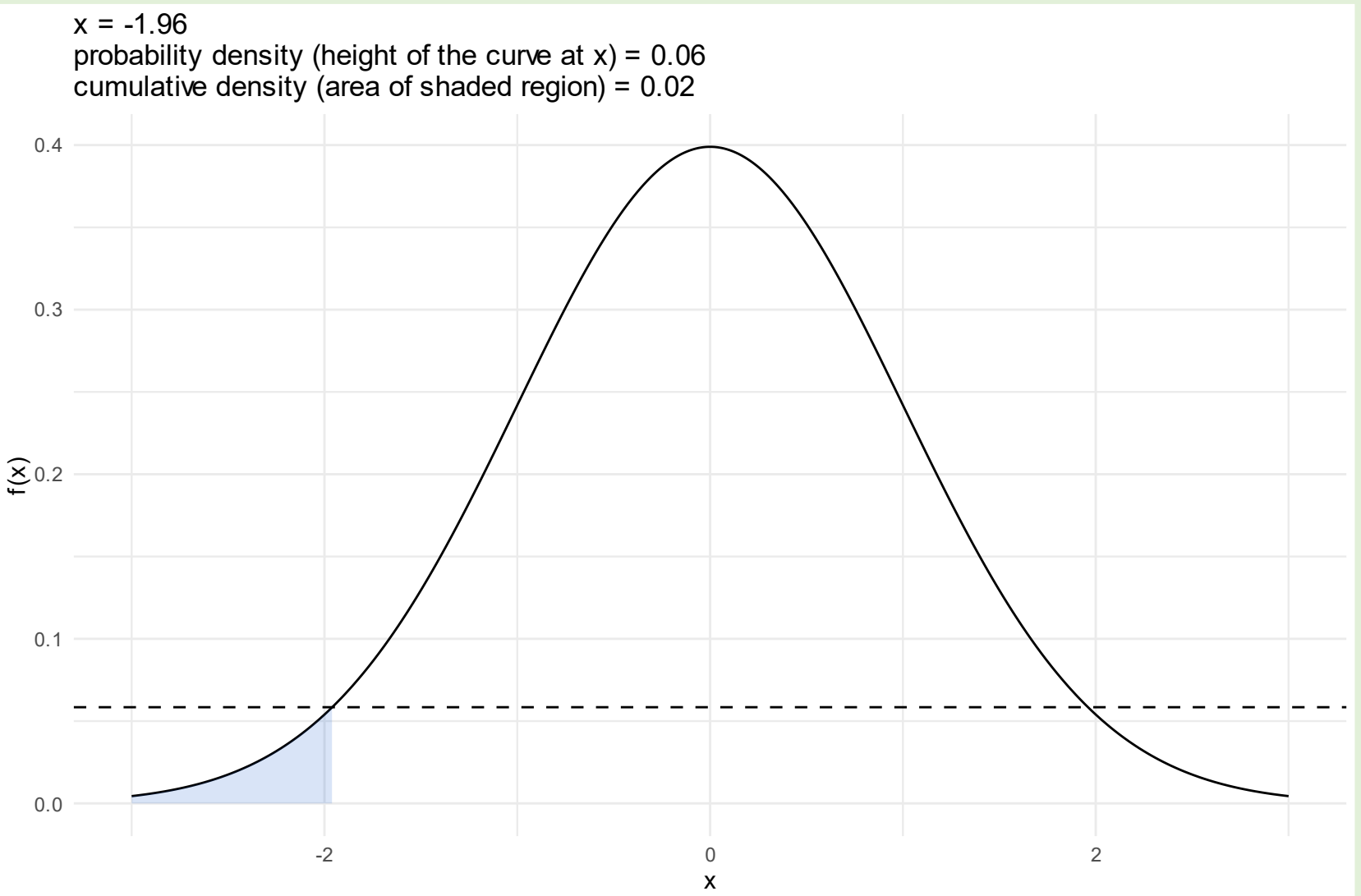
- It’s an integral!
- It is the probability of observing a value equal to or less than x.

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

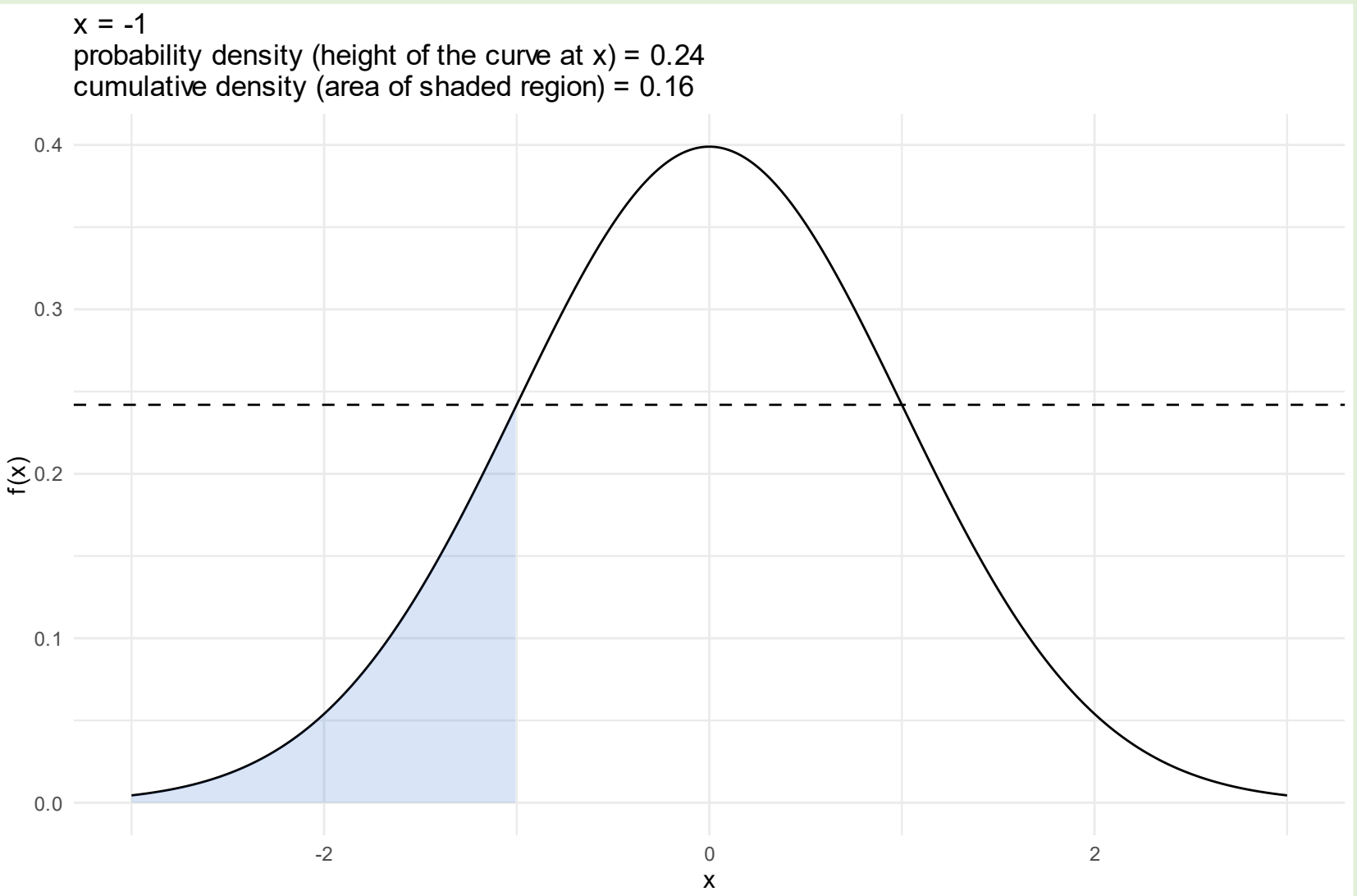


Probability Distribution Functions: Graphical Intuition

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

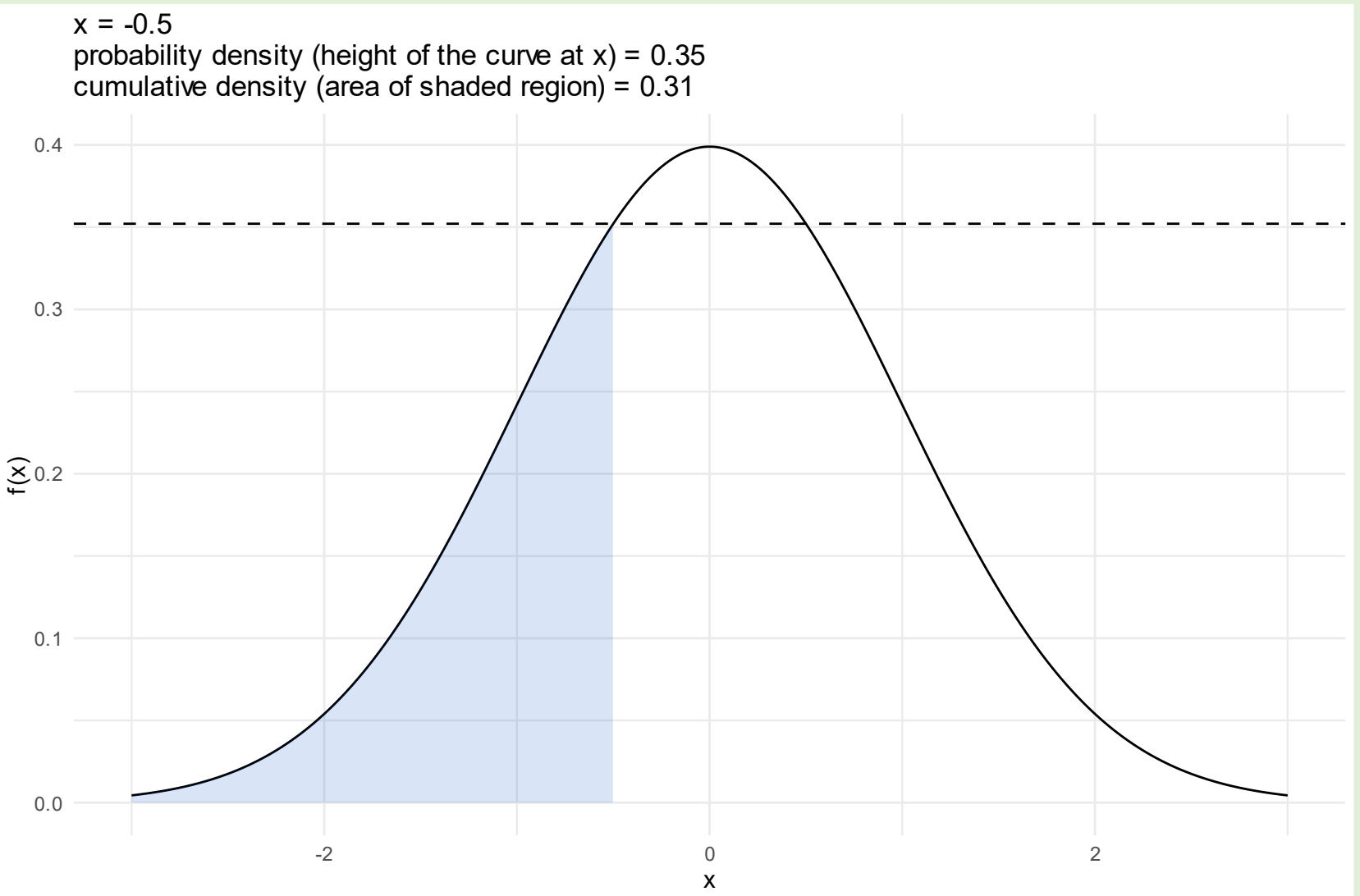


Probability Distribution Functions: Graphical Intuition

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

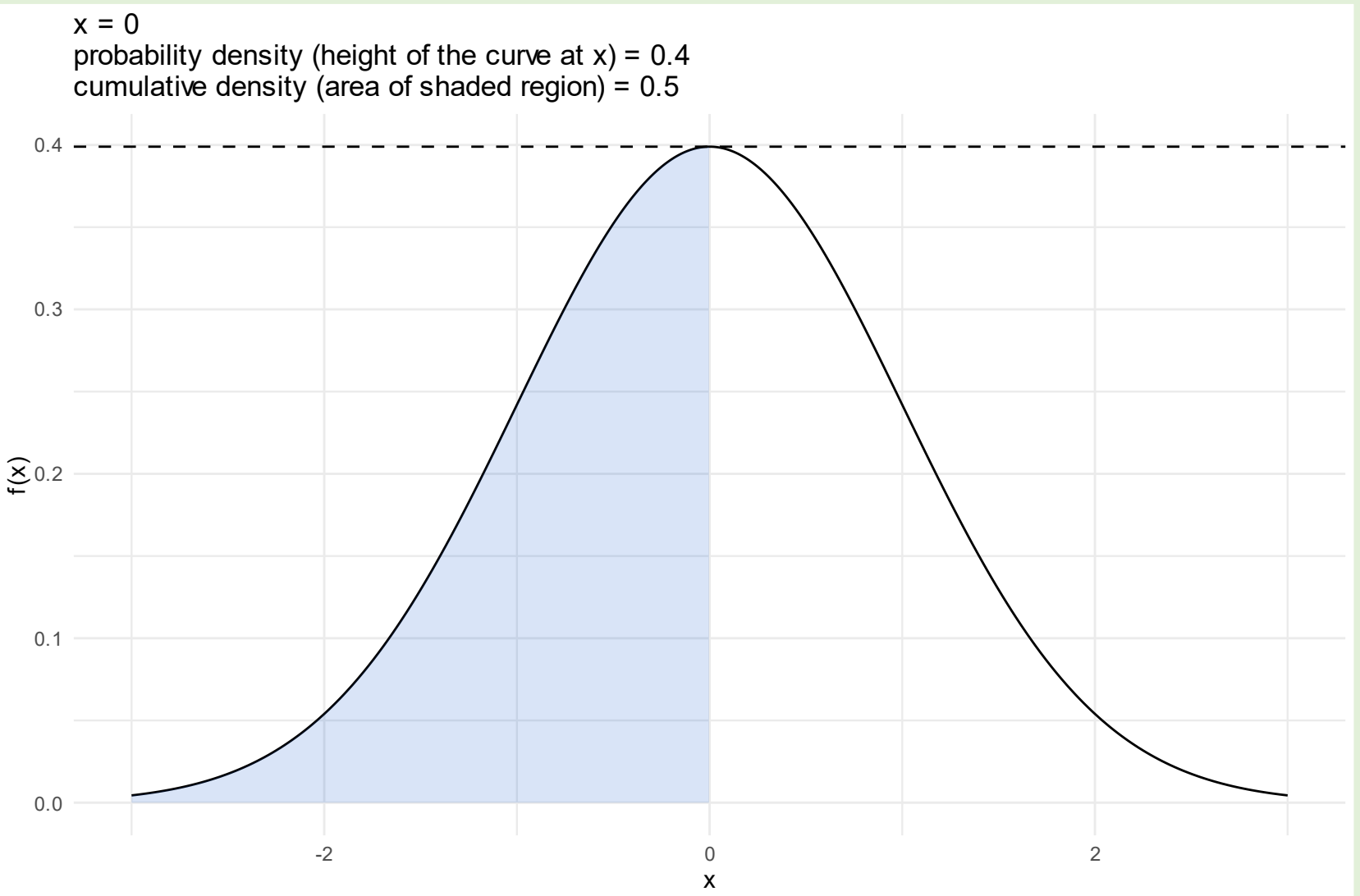


Probability Distribution Functions: Graphical Intuition

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

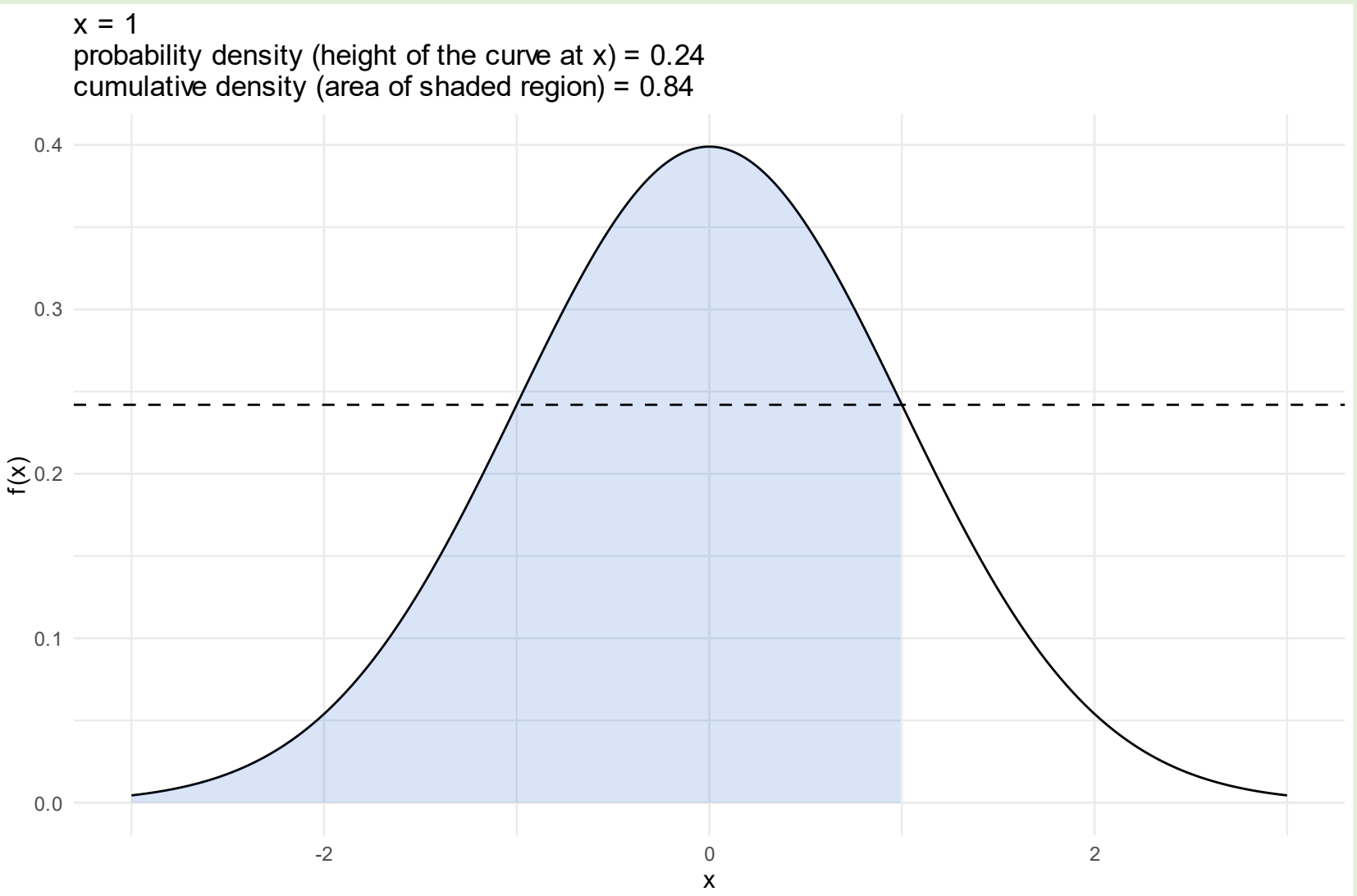


Probability Distribution Functions: Graphical Intuition

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

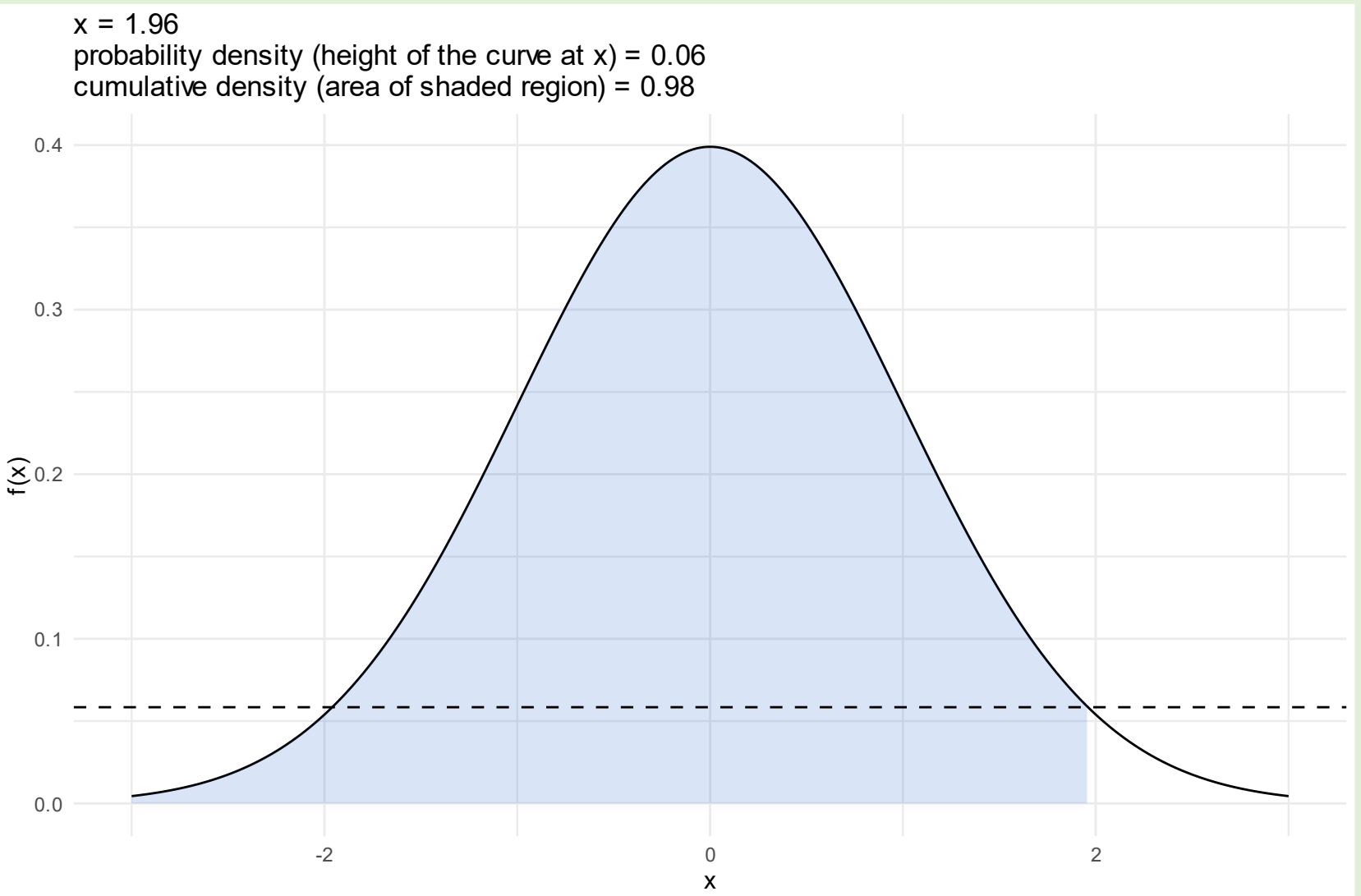


Probability Distribution Functions: Graphical Intuition

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

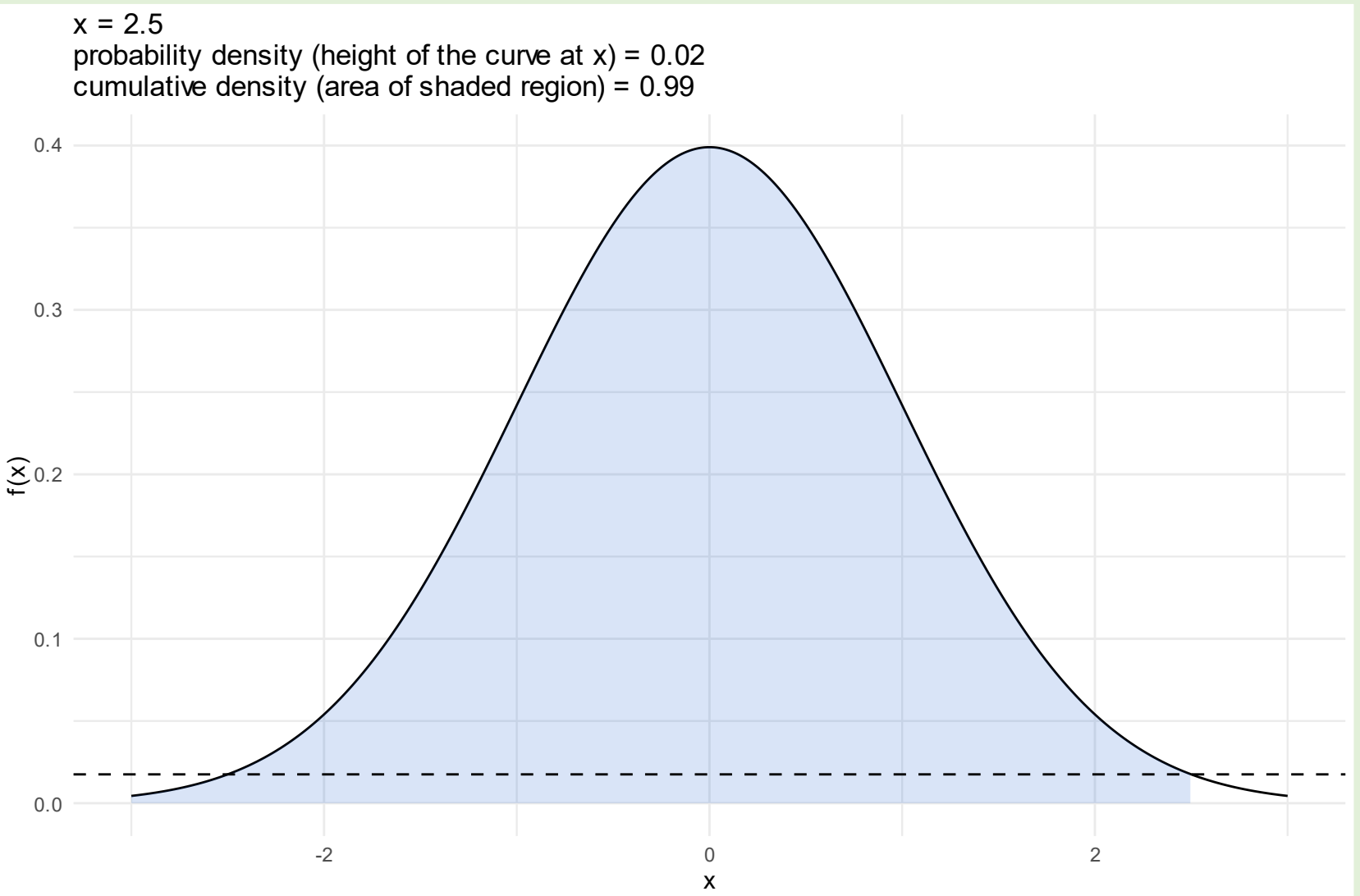


Probability Distribution Functions: Graphical Intuition

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