Introduction to Statistics and Probability

Dean Markwick

16 August 2016

Welcome!

Dean Markwick
Email: dean.markwick.15@ucl.ac.uk
Twitter: @DeanMarkwick

Course Goals

By the end of this course you will be able to understand the following:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

$$\mathrm{Var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2$$

and if a random variable x is distributed normally we can write:
$x \sim N(\mu, \sigma^2)$,
$\mathbb{E}[x] = \mu$, $\mathrm{Var}[x] = \sigma^2$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Course Plan

- Data Analysis
- Probability Theory
- Random Variables
- Basic Calculus

Small problem sheet with questions that you should attempt.

Data Analysis

- Turn data into information.

Data Analysis

- Turn data into information.
- Answer unknown questions

Data Analysis

- Turn data into information.
- Answer unknown questions
- Describe whats happening.

Data Analysis

- Turn data into information.
- Answer unknown questions
- Describe whats happening.
- Building a full statistical model

Data Analysis

- Turn data into information.
- Answer unknown questions
- Describe whats happening.
- Building a full statistical model
- Making graphs.

Data Analysis

- Turn data into information.
- Answer unknown questions
- Describe whats happening.
- Building a full statistical model
- Making graphs.
- Overall we learn something about the world that we didn't know before.

Types of Data

Data can come in many forms.

- Numerical

Types of Data

Data can come in many forms.

- Numerical
  - Temperature readings, test scores, height measurements.

Types of Data

Data can come in many forms.

- Numerical
  - Temperature readings, test scores, height measurements.
- Textual

Types of Data

Data can come in many forms.

- Numerical
  - Temperature readings, test scores, height measurements.
- Textual
  - Tweets, survey responses.

Types of Data

Data can come in many forms.

- Numerical
  - Temperature readings, test scores, height measurements.
- Textual
  - Tweets, survey responses.
- Ordinal

Types of Data

Data can come in many forms.

- Numerical
  - Temperature readings, test scores, height measurements.
- Textual
  - Tweets, survey responses.
- Ordinal
  - Customer satisfaction i.e. a number between 1 and 10.

Types of Data

Data can come in many forms.

- Numerical
  - Temperature readings, test scores, height measurements.
- Textual
  - Tweets, survey responses.
- Ordinal
  - Customer satisfaction i.e. a number between 1 and 10.
- Categorical

Types of Data

Data can come in many forms.

- Numerical
  - Temperature readings, test scores, height measurements.
- Textual
  - Tweets, survey responses.
- Ordinal
  - Customer satisfaction i.e. a number between 1 and 10.
- Categorical
  - Types of weather.

Types of Data

Data can come in many forms.

- Numerical
  - Temperature readings, test scores, height measurements.
- Textual
  - Tweets, survey responses.
- Ordinal
  - Customer satisfaction i.e. a number between 1 and 10.
- Categorical
  - Types of weather.
- Binary

Types of Data

Data can come in many forms.

- Numerical
  - Temperature readings, test scores, height measurements.
- Textual
  - Tweets, survey responses.
- Ordinal
  - Customer satisfaction i.e. a number between 1 and 10.
- Categorical
  - Types of weather.
- Binary
  - Heads or tails, dead or alive.

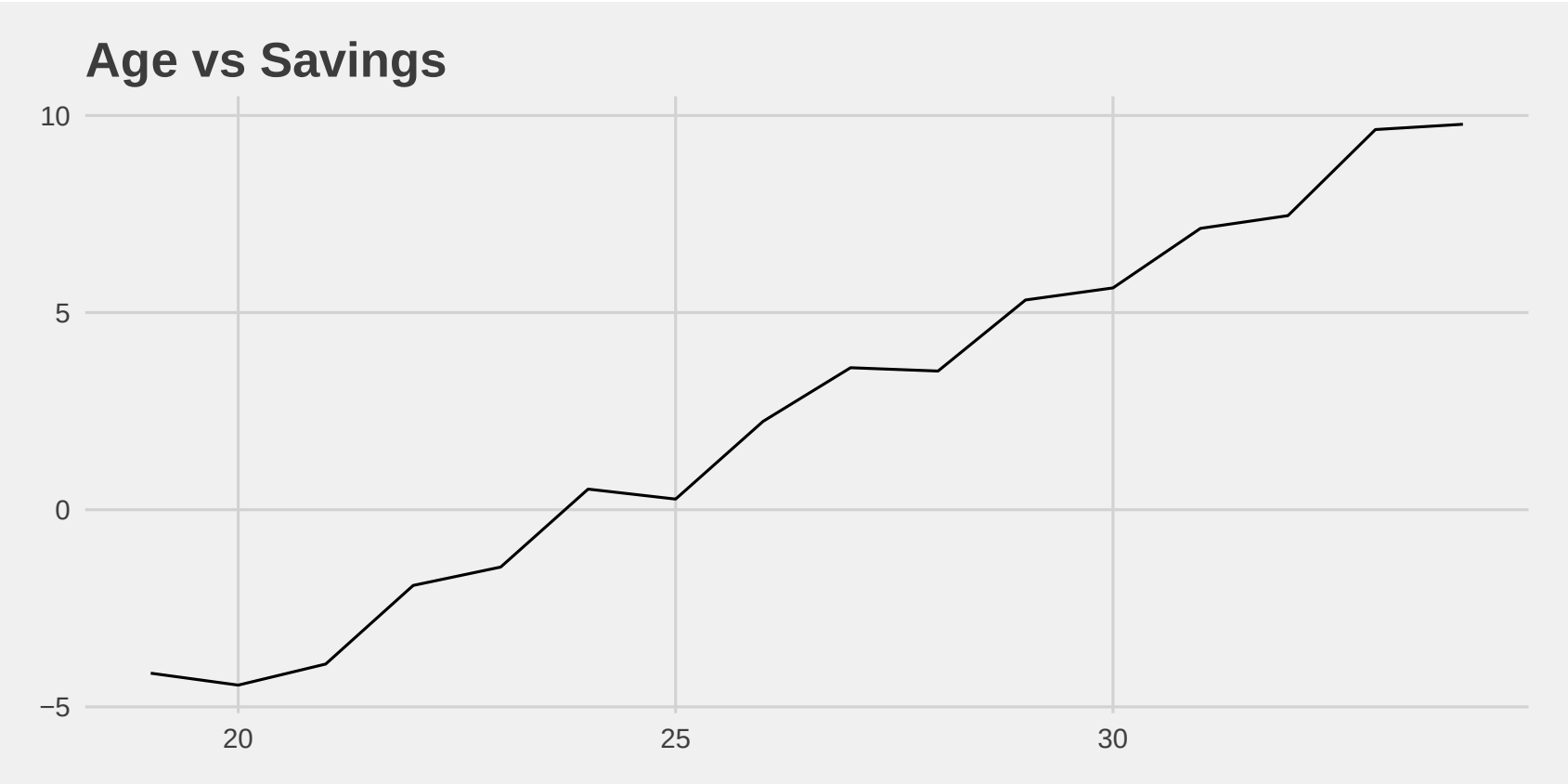
Types of Data

Data can come in many forms.

- Numerical
  - Temperature readings, test scores, height measurements.
- Textual
  - Tweets, survey responses.
- Ordinal
  - Customer satisfaction i.e. a number between 1 and 10.
- Categorical
  - Types of weather.
- Binary
  - Heads or tails, dead or alive.
- Images

Types of Data

Need to know what type of data you are dealing with so you can interpret it correctly.



In this case, negative savings = debt.

Summarising Data

Summarizing data means reducing the size.

- Averages

Summarising Data

Summarizing data means reducing the size.

- Averages
- Measuring Spread

Summarising Data

Summarizing data means reducing the size.

- Averages
- Measuring Spread
- Ranges

Summarising Data

Say we have two stocks and how much they returned every year for 5 years.

Stock1	Stock2
2	4
2	-19
1	14
1	-3
4	-13

How do we make this data more digestible?

Summarising Data: Averages

The average (commonly called the mean) is the most popular statistic to calculate.
Denoted as $\bar{x}$ (x bar).

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Add all the values together and divide by the total number of values

Summarising Data: Averages

For Stock1:

$$\bar{x} = \frac{1}{5}(2 + 2 + 1 + 1 + 4) = 2$$

For Stock2:

$$\bar{x} = \frac{1}{5}(4 + -19 + 14 + -3 + -13) = -3.4$$

Summarising Data: Measure of Spread

Measure of spread calculates how varied the data is.
One example is the **standard deviation** or **variance**.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard deviation, s, is just the square root of this.

Summarising Data: Measure of Spread

For Stock 1 the variance is:

$$s_1^2 = 1.5$$

For Stock 2:

$$s_2^2 = 173.3$$

Why are these important in understanding risk?

Summarising Data: Measure of Spread

Very general rules of thumb:

- 65-70% of the data lies within one standard deviation. The range $(\bar{x} - s, \bar{x} + s)$

But these depend on the data. It relies on the data being *normally* distributed.

Summarising Data: Measure of Spread

Very general rules of thumb:

- 65-70% of the data lies within one standard deviation. The range $(\bar{x} - s, \bar{x} + s)$
- 95% of the data lies within two standard deviations of the data. The range $(\bar{x} - 2s, \bar{x} + 2s)$

But these depend on the data. It relies on the data being *normally* distributed.

Visualising Data

Sometimes the best thing to do with data is make something pretty out of it.
Different data sets, different objectives:

- Do you want to communicate a point or just generate interest in the data?

Visualising Data

Sometimes the best thing to do with data is make something pretty out of it.
Different data sets, different objectives:

- Do you want to communicate a point or just generate interest in the data?
- What conclusions do you want the viewer to draw?

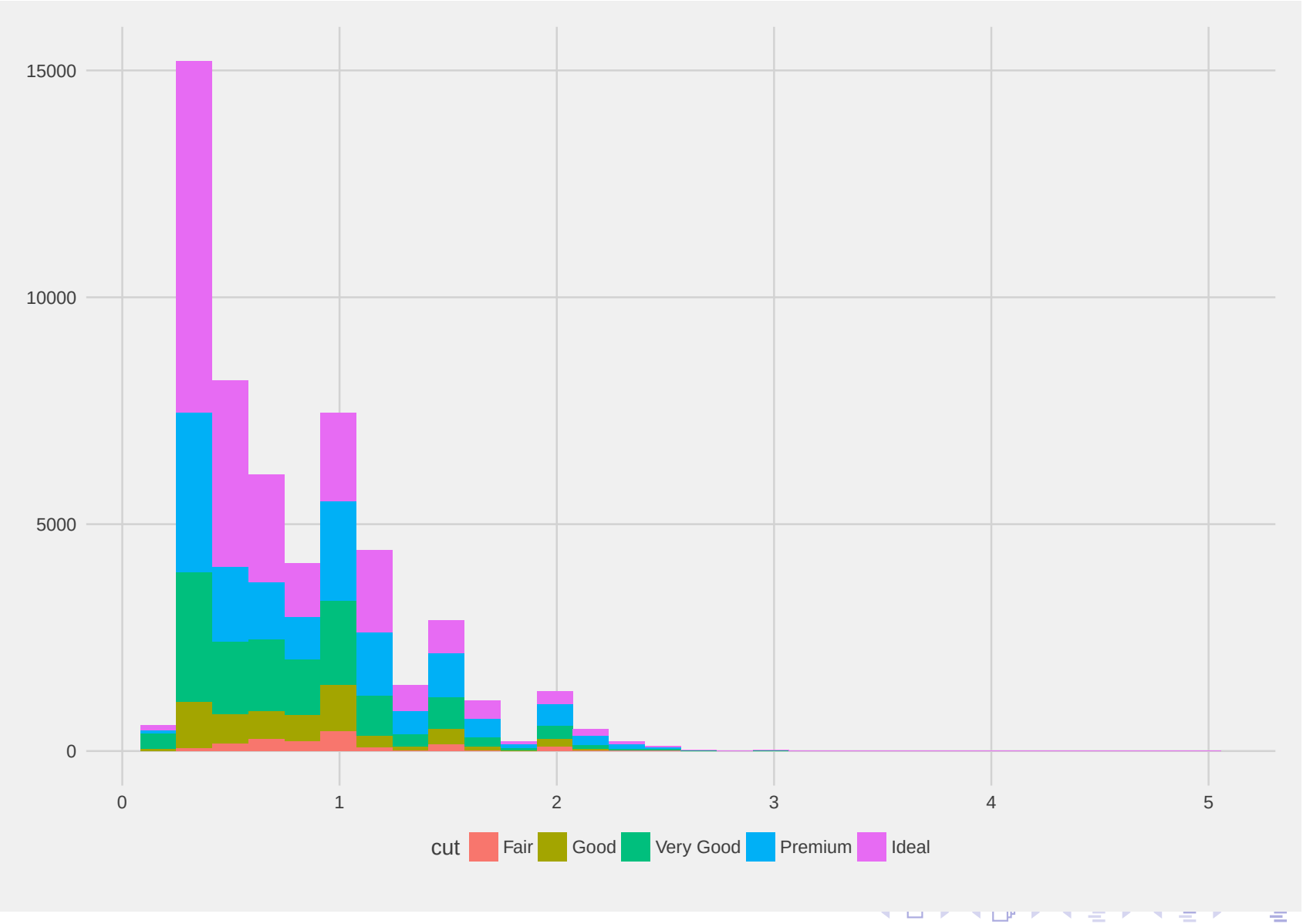
Visualising Data

Sometimes the best thing to do with data is make something pretty out of it.
Different data sets, different objectives:

- Do you want to communicate a point or just generate interest in the data?
- What conclusions do you want the viewer to draw?
- Is the message clear?

Visualising Data: Histograms

Say we had a data set on diamonds, what would we expect it to contain?



Probability

What is probability?

- Branch of mathematics trying to understand the randomness in the data generating process.

Understanding probability allows us to:

Probability

What is probability?

- Branch of mathematics trying to understand the randomness in the data generating process.
- From this we can go beyond what is in the data itself.

Understanding probability allows us to:

Probability

What is probability?

- Branch of mathematics trying to understand the randomness in the data generating process.
- From this we can go beyond what is in the data itself.

Understanding probability allows us to:

- Predict future events.

Probability

What is probability?

- Branch of mathematics trying to understand the randomness in the data generating process.
- From this we can go beyond what is in the data itself.

Understanding probability allows us to:

- Predict future events.
- Conclude about the confidence we have in the collected data.

Probability

What is probability?

- Branch of mathematics trying to understand the randomness in the data generating process.
- From this we can go beyond what is in the data itself.

Understanding probability allows us to:

- Predict future events.
- Conclude about the confidence we have in the collected data.
- Test and ask questions about the data. (Hypothesis testing)

Probability: Kolmogorov's Axioms

$$\Pr(A) \geq 0$$

$$\Pr(S) = 1$$

$$\Pr(A \text{or} B) = \Pr(A) + \Pr(B)$$

In words:

- The probability of A happening is greater or equal to 0.

Probability: Kolmogorov's Axioms

$$\Pr(A) \geq 0$$

$$\Pr(S) = 1$$

$$\Pr(A \text{or} B) = \Pr(A) + \Pr(B)$$

In words:

- The probability of A happening is greater or equal to 0.
- The probability of *something* happening is 1.

Probability: Kolmogorov's Axioms

$$\Pr(A) \geq 0$$

$$\Pr(S) = 1$$

$$\Pr(A \text{or} B) = \Pr(A) + \Pr(B)$$

In words:

- The probability of A happening is greater or equal to 0.
- The probability of *something* happening is 1.
- If A and B cannot both happen then the probability that at least A or B happens is the probability of them individually happening added together.

Probability: Dice roll

Consider a six sided die.

- What is the sample space, S?

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$\Pr(1) = \frac{1}{6}$$

Probability: Dice roll

Consider a six sided die.

- What is the sample space, S?
- What is the probability of a 1 being thrown?

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$\Pr(1) = \frac{1}{6}$$

Probability: Dice roll

What about the probability that a roll is 2 or less?

$$A = \{1, 2\}$$

$$\Pr(A) = \Pr(1) + \Pr(2)$$

$$\Pr(A) = \frac{1}{3}$$

Conditional Probability

When we know one thing, can we make a better guess for another thing?

- Can your performance at school predict how well you do at University?

Conditional Probability

When we know one thing, can we make a better guess for another thing?

- Can your performance at school predict how well you do at University?
- Does the maximum height of tidal waves over the last 100 years help us predict the worst case scenario for the next hundred years?

Conditional Probability

When we know one thing, can we make a better guess for another thing?

- Can your performance at school predict how well you do at University?
- Does the maximum height of tidal waves over the last 100 years help us predict the worst case scenario for the next hundred years?
- Does how much money a Premier League football team spends in the transfer market help predict how well they will do?

Conditional Probability

All these questions can be formulated in the same way.
The probability A occurs ***given*** *that we know B has occured*.
Mathematically

$$\Pr(A|B) = \frac{\Pr(A \text{and} B)}{\Pr(B)}$$

But this can be understood using Venn diagrams.

Conditional Probability: Venn Diagrams

Conditional Probability: Independent Events

Events are independent if knowledge of one doesn't effect the other.

$$\Pr(A|B) = \Pr(A)$$

For example, roll one die then a different die. The two outcomes are independent of each other.

Random Variables

So with the basics in probability we can now look at modelling real world events.
But what is a random variable?

- A number that varies randomly.

Random Variables

So with the basics in probability we can now look at modelling real world events.
But what is a random variable?

- A number that varies randomly.
- The outcome of an experiment before the experiment has happened.

Random Variables

So with the basics in probability we can now look at modelling real world events.
But what is a random variable?

- A number that varies randomly.
- The outcome of an experiment before the experiment has happened.
- A map from the sample space S into the set of numbers.

Discrete Random Variables

These take on a finite number of values. The simplest example is the *binary random variable*:

$$X = \begin{cases} 1 & \text{Heads} \\ 0 & \text{Tails} \end{cases}$$

X is the random variable.
It is the result of the coin flip. It maps the result (heads or tails) to a number.
The key result is that X is unknown before we flip the coin.

Probability Distributions

So how can we model X?
As it is a coin flip there are two outcomes with equal chance.

$$\Pr(X = 1) = \frac{1}{2}$$

$$\Pr(X = 0) = \frac{1}{2}$$

So now we have a probability distribution of X.

Probability Distributions: Two Coin Flips

What is the sample space if we flip a coin twice?
If X is the number of heads:

$$\Pr(X = 0) = \frac{1}{4}$$

$$\Pr(X = 1) = \frac{1}{2}$$

$$\Pr(X = 2) = \frac{1}{4}$$

Do the three axioms hold?

Probability Distributions

How many years will it be until the next flood which is severe enough to break the Thames barrier?
The simple model:

- The probability of a bad flood in a single year is p

$$\Pr(N = n) = (1 - p)^{n-1} p$$

Where N is the number of years until a bad flood.
Is this a reasonable model?
How do we communicate this model?

Probability Distributions

How many years will it be until the next flood which is severe enough to break the Thames barrier?
The simple model:

- The probability of a bad flood in a single year is p
- Assume that each year is independent.

$$\Pr(N = n) = (1 - p)^{n-1} p$$

Where N is the number of years until a bad flood.
Is this a reasonable model?
How do we communicate this model?

The Expected Value

An idealized average of a random variable.
In general we write:

$$\mathbb{E}[X] = \sum_x xP(X = x)$$

Sum over all the possible values of x.

$$xP(X = x)$$

weights each value by the probability it will occur.

The Expected Value

What's the expected value of a dice role?

The Expected Value: Why?

Can be used to calculate probability for gambling games.

$$\mathbb{E}\left[\text{Profit}\right] = \text{winnings} \times P(\text{win}) + \text{losses} \times P(\text{loss})$$

Should you play the national lottery?

Continuous Data

Most things in life don't take on a discrete set of values.

- Height

These are continuous random variables and need a require a bit more theory.

Continuous Data

Most things in life don't take on a discrete set of values.

- Height
- Distance etc.

These are continuous random variables and need a require a bit more theory.

Continuous Random Variable

If a random variable is continuous it has a probability density function $f(x)$.
The probability of X being within an interval:

$$P(a < X < b) = \int_a^b f(x)\mathrm{d}x$$

and $f(x)$ satisfies the following conditions

- $f(x) \geq 0$

Continuous Random Variable

If a random variable is continuous it has a probability density function $f(x)$.
The probability of X being within an interval:

$$P(a < X < b) = \int_a^b f(x)\mathrm{d}x$$

and $f(x)$ satisfies the following conditions

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$

Continuous Random Variable

So what do these symbols mean?

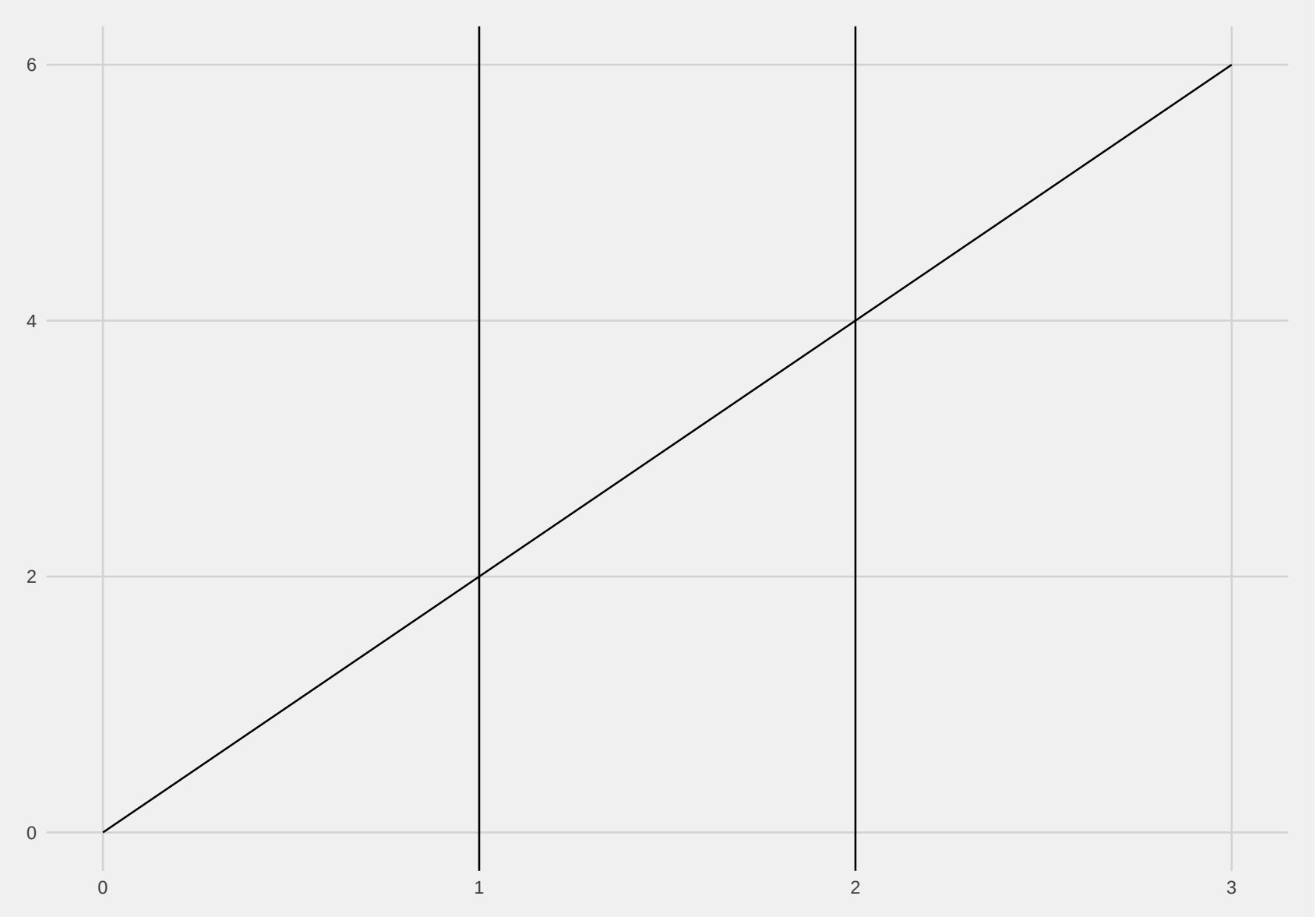
$$\int_a^b f(x)\mathrm{d}x$$

Calculate the area under the curve $f(x)$ between a and b!

Integration

Lets say $f(x) = 2x$.

What is the area under the curve between 1 and 2?



Integration

Integration is just the limiting process, where we us infinitely many rectangles to calculate the area. A first course in calculus goes through this more rigorously. But for now, all you need to remember is that

$$\int_a^b f(x)\mathrm{d}x$$

is the area of $f(x)$ between a and b.

Expectation and Varaince

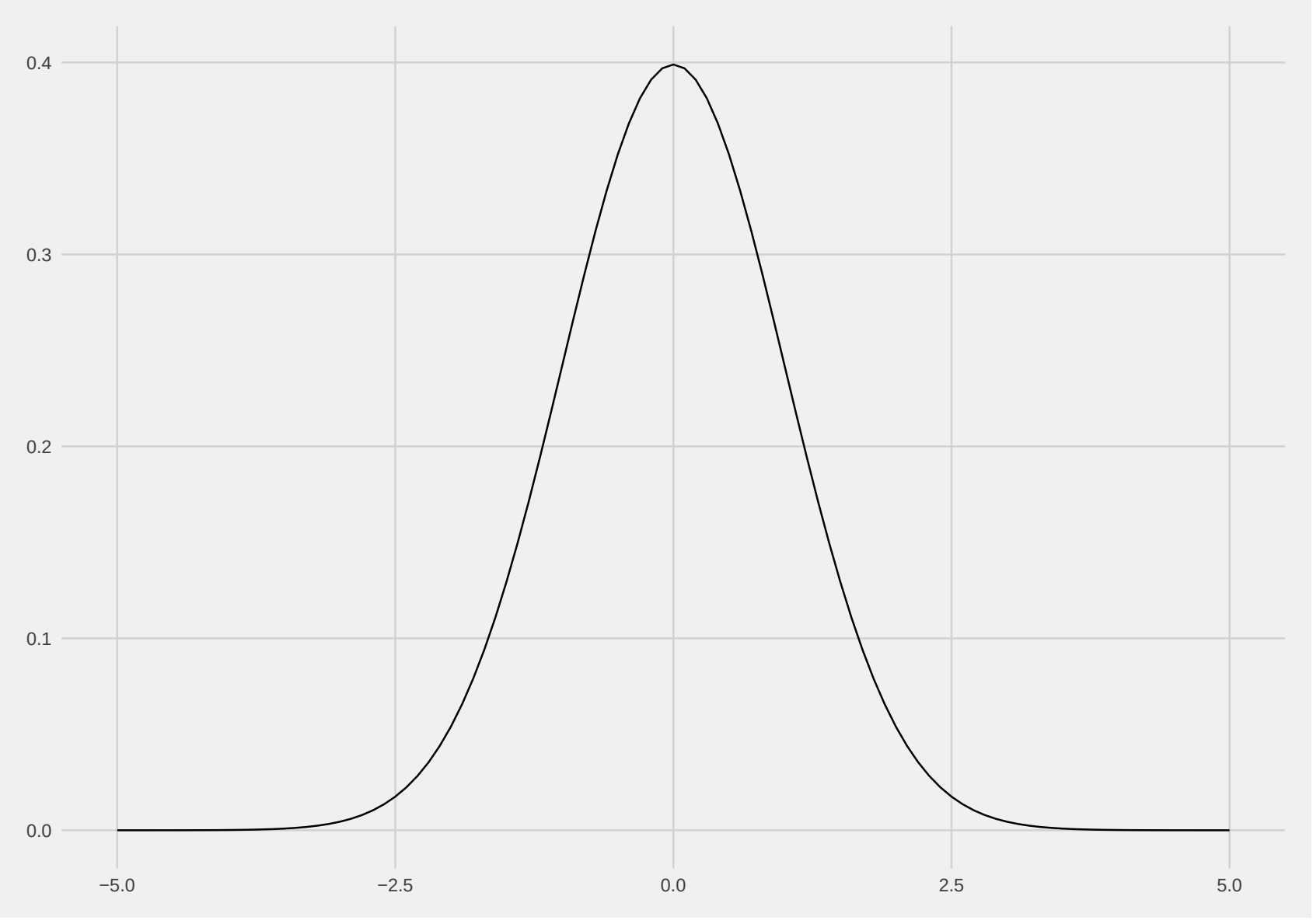
Now we now how to calculate the probability that a continuous variable is equal to some value, how do we calculate the expectation?
We replace the sums with integrals.

$$\mathbb{E}[X] = \int x f(x) \mathrm{d}x$$

$$\mathrm{Var}[X] = \int (x - \mathbb{E}[X])^2 f(x) \mathrm{d}x$$

The Normal Distribution

Most things in life are normally distributed.



This is the probability density of the normal distribution.

The Normal Distribution

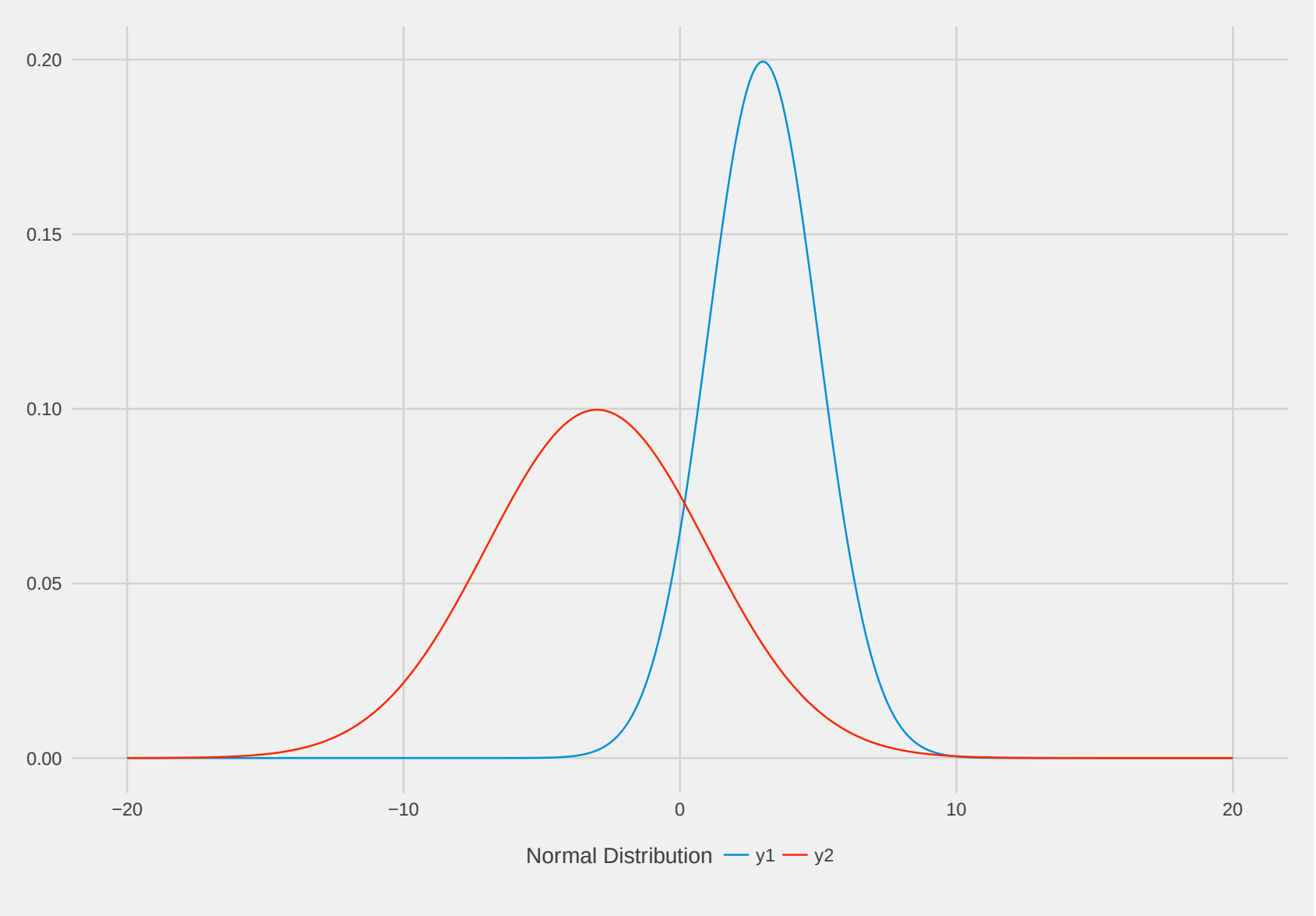
The normal distribution is characterized by two parameters. The mean and variance. So if X is normally distributed we write

$$X \sim N(\mu, \sigma^2)$$

where μ is the mean and σ^2 is the variance.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

The Normal Distribution



$N(3, 4)$ or $N(-3, 16)$