Introduction to Data Science - 1MS041

Benny Avelin
Department of Mathematics

HT 2023

So our story begins

We have seen the problems, now lets start with the language.

- Experiment: is an activity or procedure that produces distinct, well-defined possibilities called **outcomes**.

So our story begins

We have seen the problems, now lets start with the language.

- Experiment: is an activity or procedure that produces distinct, well-defined possibilities called **outcomes**.
- The set of all outcomes is called the **sample space**, and is denoted by Ω.

So our story begins

We have seen the problems, now lets start with the language.

- Experiment: is an activity or procedure that produces distinct, well-defined possibilities called **outcomes**.
- The set of all outcomes is called the **sample space**, and is denoted by Ω.
- Trial: doing the experiment once and getting an outcome.

So our story begins

We have seen the problems, now lets start with the language.

- Experiment: is an activity or procedure that produces distinct, well-defined possibilities called **outcomes**.
- The set of all outcomes is called the **sample space**, and is denoted by Ω.
- Trial: doing the experiment once and getting an outcome.
- The subsets of Ω are called **events**.

So our story begins

We have seen the problems, now lets start with the language.

- Experiment: is an activity or procedure that produces distinct, well-defined possibilities called **outcomes**.
- The set of all outcomes is called the **sample space**, and is denoted by Ω.
- Trial: doing the experiment once and getting an outcome.
- The subsets of Ω are called **events**.
- Given an outcome $\omega \in \Omega$ we say that the event $E \subset \Omega$ **occured** if $\omega \in E$.

Some standard examples of experiments are the following:

1. $\Omega = \{\text{Defective, Non-defective}\}$ if our experiment is to inspect a light bulb.
   There are only two outcomes here, so $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \text{Defective}$ and $\omega_2 = \text{Non-defective}$.

Some standard examples of experiments are the following:

1. $\Omega = \{\text{Defective, Non-defective}\}$ if our experiment is to inspect a light bulb.
   There are only two outcomes here, so $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \text{Defective}$ and $\omega_2 = \text{Non-defective}$.
2. $\Omega = \{\text{Heads, Tails}\}$ if our experiment is to note the outcome of a coin toss.
   This time, $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \text{Heads}$ and $\omega_2 = \text{Tails}$.

Some standard examples of experiments are the following:

1. $\Omega = \{\text{Defective, Non-defective}\}$ if our experiment is to inspect a light bulb.
   There are only two outcomes here, so $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \text{Defective}$ and $\omega_2 = \text{Non-defective}$.
2. $\Omega = \{\text{Heads, Tails}\}$ if our experiment is to note the outcome of a coin toss.
   This time, $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \text{Heads}$ and $\omega_2 = \text{Tails}$.
3. If our experiment is to roll a die then there are six outcomes corresponding to the number that shows on the top. For this experiment, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
   Some examples of events are the set of odd numbered outcomes $A = \{1, 3, 5\}$, and the set of even numbered outcomes $B = \{2, 4, 6\}$.
   The simple events of Ω are $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$, and $\{6\}$.

The long-term relative frequency (LTRF) idea

Suppose we are interested in the fairness of a coin, i.e. if landing Heads has the same "probability" as landing Tails.

- Toss it n times and call $N(\{\mathtt{H}\}, \mathtt{n})$ the fraction of times we observed Heads out of n tosses.

The long-term relative frequency (LTRF) idea

Suppose we are interested in the fairness of a coin, i.e. if landing Heads has the same "probability" as landing Tails.

- Toss it n times and call $N(\{\mathtt{H}\}, \mathtt{n})$ the fraction of times we observed Heads out of n tosses.
- We expect that if a coin is fair, then roughly half the times we should see Head and the other half Tails. Thus we expect that $N(\{\mathtt{H}\}, \mathtt{n})$ should as n is very large, be close to 0.5.

The long-term relative frequency (LTRF) idea

Suppose we are interested in the fairness of a coin, i.e. if landing Heads has the same “probability” as landing Tails.

- Toss it n times and call $N(\{\mathtt{H}\}, \mathtt{n})$ the fraction of times we observed Heads out of n tosses.
- We expect that if a coin is fair, then roughly half the times we should see Head and the other half Tails. Thus we expect that $N(\{\mathtt{H}\}, \mathtt{n})$ should as n is very large, be close to 0.5.
- If it is not, then intuitively we would say that it is not a fair coin.

Rules of the game

1. **Something Happens**:

$$N(\{\mathtt{H}\} \cup \{\mathtt{T}\}, \mathtt{n}) = \frac{\mathtt{n}}{\mathtt{n}} = 1.$$

Rules of the game

1. **Something Happens**:

$$N(\{\mathtt{H}\} \cup \{\mathtt{T}\}, \mathtt{n}) = \frac{\mathtt{n}}{\mathtt{n}} = \mathtt{1}.$$

2. **Addition Rule**: Mutually exclusive events are additive

$$N(\{\mathtt{H}\} \cup \{\mathtt{T}\}, \mathtt{n}) = \mathtt{N}(\{\mathtt{H}\}, \mathtt{n}) + \mathtt{N}(\{\mathtt{T}\}, \mathtt{n}).$$

Rules of the game

1. **Something Happens**:

$$N(\{\mathtt{H}\} \cup \{\mathtt{T}\}, \mathtt{n}) = \frac{\mathtt{n}}{\mathtt{n}} = 1.$$

2. **Addition Rule**: Mutually exclusive events are additive

$$N(\{\mathtt{H}\} \cup \{\mathtt{T}\}, \mathtt{n}) = \mathtt{N}(\{\mathtt{H}\}, \mathtt{n}) + \mathtt{N}(\{\mathtt{T}\}, \mathtt{n}).$$

3. **Independence**: The outcome of any individual coin-toss does not affect that of another.

Formalisation of these concepts

We saw that if we have events, we can take their union and it is still fine. We have the following rules

Definition

Let Ω be a set: We say that a collection of subsets of Ω, $\mathcal{F}$ is a **sigma-algebra**/ **sigma-field**/ σ-**algebra** if it satisfies the following properties:

1. $\mathcal{F}$ contains Ω, i.e. $\Omega \in \mathcal{F}$.
2. The collection $\mathcal{F}$ is closed under complementation

$$A \in \mathcal{F} \implies A^C \in \mathcal{F}.$$

3. The collection $\mathcal{F}$ is closed under countable unions

$$A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_i A_i \in \mathcal{F}.$$

Probability

Definition

Let us have an experiment with sample space Ω. Let $\mathcal{F}$ denote σ-algebra. A **probability measure** is a function $\mathbb{P} : \mathcal{F} \to [0,1]$ satisfying the following conditions:

1. The 'Something Happens' axiom holds, i.e. $\mathbb{P}(\Omega) = 1$.
2. The 'Addition Rule' axiom holds, i.e. for $A, B \in \mathcal{F}$:

$$A \cap B = \emptyset \quad \Longrightarrow \quad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \ .$$

We call elements of $\mathcal{F}$, events and we will call $(\Omega, \mathcal{F}, \mathbb{P})$ a **probability triple**.

Interpretation

How do we interpret the concept of σ-algebra?

- It is the set of events that we can prescribe a probability to...

Interpretation

How do we interpret the concept of σ-algebra?

- It is the set of events that we can prescribe a probability to...
- I.e. an event E is in $\mathcal{F}$ if $N(E, n)$ makes sense, that is, if we can observe if E happened or not.

Interpretation

How do we interpret the concept of σ-algebra?

- It is the set of events that we can prescribe a probability to...
- I.e. an event E is in $\mathcal{F}$ if $N(E, n)$ makes sense, that is, if we can observe if E happened or not.
- Consider the coin toss, then everything was fine, we could see Heads and we could see Tails, no problem.

Interpretation

How do we interpret the concept of σ-algebra?

- It is the set of events that we can prescribe a probability to...
- I.e. an event E is in $\mathcal{F}$ if $N(E, n)$ makes sense, that is, if we can observe if E happened or not.
- Consider the coin toss, then everything was fine, we could see Heads and we could see Tails, no problem.

Now consider the following problem:

Example

Suppose that you toss a coin at the same time that your friend tosses another coin in the building next doors. You cannot see your friends coin, but it got flipped nontheless. The sample space is $\Omega = \{\mathtt{HH}, \mathtt{TH}, \mathtt{HT}, \mathtt{TT}\}$, but which events can we observe has happened or not?

Let the first H/T be yours and the second be your friends, which ones are observable for you?

1. {HH}?

Let the first H/T be yours and the second be your friends, which ones are observable for you?

1. $\{\texttt{HH}\}$?
2. $\{\texttt{HT}\}$?

Let the first `H`/`T` be yours and the second be your friends, which ones are observable for you?

1. $\{\texttt{HH}\}$?
2. $\{\texttt{HT}\}$?
3. $\{\texttt{TT}\}$?

Let the first $\mathtt{H}/\mathtt{T}$ be yours and the second be your friends, which ones are observable for you?

1. $\{\mathtt{HH}\}$?
2. $\{\mathtt{HT}\}$?
3. $\{\mathtt{TT}\}$?
4. $\{\mathtt{HH}, \mathtt{HT}\}$?

Let the first $\mathtt{H}/\mathtt{T}$ be yours and the second be your friends, which ones are observable for you?

1. $\{\mathtt{HH}\}$?
2. $\{\mathtt{HT}\}$?
3. $\{\mathtt{TT}\}$?
4. $\{\mathtt{HH}, \mathtt{HT}\}$?
5. $\{\mathtt{TH}, \mathtt{TT}\}$?

Let the first $\mathtt{H}/\mathtt{T}$ be yours and the second be your friends, which ones are observable for you?

1. $\{\mathtt{HH}\}$?
2. $\{\mathtt{HT}\}$?
3. $\{\mathtt{TT}\}$?
4. $\{\mathtt{HH}, \mathtt{HT}\}$?
5. $\{\mathtt{TH}, \mathtt{TT}\}$?

Note: we could in this case completely ignore what our friend is up to and re-define $\Omega = \{\mathtt{H}, \mathtt{T}\}$.

Conditional probability

Conditional probabilities are often expressed in English by phrases such as:

- "If A happens, what is the probability that B happens?"
- "What is the probability that A happens if B happens?"
- "Given that B occurs what is the probability that A occurs?"

Conditional probability

Conditional probabilities are often expressed in English by phrases such as:

- "If A happens, what is the probability that B happens?"
- "What is the probability that A happens if B happens?"
- "Given that B occurs what is the probability that A occurs?"

Definition

Consider a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Let $A, B \in \mathcal{F}$ (events), such that $\mathbb{P}(A) \neq 0$. Then, we define the **conditional probability** of B given A by,

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \ .$$

Model problem: SMS spam filtering

Recall the SMS spam filter problem, i.e. you would like to filter your SMS (or similar instant messaging) texts as "Spam" or not.
We asked the following questions

1. What is the probability that an incoming text is spam?
2. Given that you see the word "free" in the text, what is now the probability that it is spam?

What is the probability that it is spam?

1. Experiment:

What is the probability that it is spam?

1. Experiment: Recieving SMS texts and classifying them as spam / not spam

What is the probability that it is spam?

1. Experiment: Recieving SMS texts and classifying them as spam / not spam
2. We do a single trial: A recieved SMS text which contained is a spam.

What is the probability that it is spam?

1. Experiment: Recieving SMS texts and classifying them as spam / not spam
2. We do a single trial: A recieved SMS text which contained is a spam.
3. The sample space is $\Omega = \{"\texttt{spam}", "\texttt{not spam}"\}$

What is the probability that it is spam?

1. The sample space is $\Omega = \{\text{"}\texttt{spam}\text{"}, \text{"}\texttt{not spam}\text{"}\}$.
2. What is the probability that an incoming text is spam? Let the event $A =$

What is the probability that it is spam?

1. The sample space is $\Omega = \{''\texttt{spam}'', ''\texttt{not spam}''\}$.
2. What is the probability that an incoming text is spam? Let the event $A = \{''spam''\}$,

What is the probability that it is spam?

1. The sample space is $\Omega = \{"\texttt{spam}", "\texttt{not spam}"\}$.
2. What is the probability that an incoming text is spam? Let the event $A = \{"spam"\}$, then the answer is

$$\mathbb{P}(A)$$

What is the probability that it is spam?

1. The sample space is $\Omega = \{\text{"}\texttt{spam}\text{"}, \text{"}\texttt{not spam}\text{"}\}$.
2. What is the probability that an incoming text is spam? Let the event $A = \{\text{"}spam\text{"}\}$, then the answer is

$$\mathbb{P}(A)$$

3. For our example with only one trial, we get

$$N(A, 1)$$

What is the probability that it is spam?

1. The sample space is $\Omega = \{"\texttt{spam}", "\texttt{not spam}"\}$.
2. What is the probability that an incoming text is spam? Let the event $A = \{"spam"\}$, then the answer is

$$\mathbb{P}(A)$$

3. For our example with only one trial, we get

$$N(A, 1) = 1.$$

Given that you see the word "free" in the text, what is now the probability that it is spam?

1. Experiment:

Given that you see the word "free" in the text, what is now the probability that it is spam?

1. Experiment: Recieving SMS texts checking if "free" is in the text and classifying them as spam / not spam

Given that you see the word "free" in the text, what is now the probability that it is spam?

1. Experiment: Recieving SMS texts checking if "free" is in the text and classifying them as spam / not spam
2. We do a single trial: A recieved SMS text which contained the word "free" but was not a spam.

Given that you see the word "free" in the text, what is now the probability that it is spam?

1. Experiment: Recieving SMS texts checking if "free" is in the text and classifying them as spam / not spam
2. We do a single trial: A recieved SMS text which contained the word "free" but was not a spam.
3. The sample space is $\Omega = \{\text{"}\texttt{free, spam}\text{"}, \text{"}\texttt{no free, spam}\text{"}, \text{"}\texttt{free, not spam}\text{"}, \text{"}\texttt{no free, not spam}\text{"}\}$.

1. The sample space is $\Omega = \{$"$\texttt{free, spam}$", "$\texttt{no free, spam}$", "$\texttt{free, not spam}$", "$\texttt{no free, not spam}$"$\}$.

1. The sample space is $\Omega = \{"\texttt{free}, \texttt{spam}", "\texttt{no free}, \texttt{spam}", "\texttt{free}, \texttt{not spam}", "\texttt{no free}, \texttt{not spam}"\}$.
2. What is the probability that an incoming text is spam given that it contains free? The event that free appeared is the outcomes $B =$

1. The sample space is $\Omega = \{\text{"}\texttt{free, spam}\text{"}, \text{"}\texttt{no free, spam}\text{"}, \text{"}\texttt{free, not spam}\text{"}, \text{"}\texttt{no free, not spam}\text{"}\}$.
2. What is the probability that an incoming text is spam given that it contains free? The event that free appeared is the outcomes $B = \{\text{"}free, spam\text{"}, \text{"}free, not\ spam\text{"}\}$, and the event that it is spam is the event $A =$

1. The sample space is $\Omega = \{\text{"}\texttt{free}, \texttt{spam}\text{"}, \text{"}\texttt{no free}, \texttt{spam}\text{"}, \text{"}\texttt{free}, \texttt{not spam}\text{"}, \text{"}\texttt{no free}, \texttt{not spam}\text{"}\}$.
2. What is the probability that an incoming text is spam given that it contains free? The event that free appeared is the outcomes $B = \{\text{"}free, spam\text{"}, \text{"}free, not\ spam\text{"}\}$, and the event that it is spam is the event $A = \{\text{"}free, spam\text{"}, \text{"}no\ free, spam\text{"}\}$

   $$\mathbb{P}(A \mid B) =$$
   $$\mathbb{P}(\{\text{"}free, spam\text{"}, \text{"}no\ free, spam\text{"}\} \mid \{\text{"}free, spam\text{"}, \text{"}free, not\ spam\text{"}\})$$

1. The sample space is $\Omega = \{''\texttt{free}, \texttt{spam}'', ''\texttt{no free}, \texttt{spam}'', ''\texttt{free}, \texttt{not spam}'', ''\texttt{no free}, \texttt{not spam}''\}$.
2. What is the probability that an incoming text is spam given that it contains free? The event that free appeared is the outcomes $B = \{''\mathit{free}, \mathit{spam}'', ''\mathit{free}, \mathit{not\ spam}''\}$, and the event that it is spam is the event $A = \{''\mathit{free}, \mathit{spam}'', ''\mathit{no\ free}, \mathit{spam}''\}$

   $\mathbb{P}(A \mid B) =$
   $\mathbb{P}(\{''\mathit{free}, \mathit{spam}'', ''\mathit{no\ free}, \mathit{spam}''\} \mid \{''\mathit{free}, \mathit{spam}'', ''\mathit{free}, \mathit{not\ spam}''\})$

3. Which by definition becomes

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{''\mathit{free}, \mathit{spam}''\})}{\mathbb{P}(\{''\mathit{free}, \mathit{spam}'', ''\mathit{free}, \mathit{not\ spam}''\})}$$

1. The sample space is $\Omega = \{\text{"}\texttt{free}, \texttt{spam}\text{"}, \text{"}\texttt{no free}, \texttt{spam}\text{"}, \text{"}\texttt{free}, \texttt{not spam}\text{"}, \text{"}\texttt{no free}, \texttt{not spam}\text{"}\}$.
2. What is the probability that an incoming text is spam given that it contains free? The event that free appeared is the outcomes $B = \{\text{"}free, spam\text{"}, \text{"}free, not\ spam\text{"}\}$, and the event that it is spam is the event $A = \{\text{"}free, spam\text{"}, \text{"}no\ free, spam\text{"}\}$

   $$\mathbb{P}(A \mid B) =$$
   $$\mathbb{P}(\{\text{"}free, spam\text{"}, \text{"}no\ free, spam\text{"}\} \mid \{\text{"}free, spam\text{"}, \text{"}free, not\ spam\text{"}\})$$

3. Which by definition becomes

   $$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{\text{"}free, spam\text{"}\})}{\mathbb{P}(\{\text{"}free, spam\text{"}, \text{"}free, not\ spam\text{"}\})}$$

4. This becomes for our example $N(\{\text{"}free, spam\text{"}\}, 1)/N(\{\text{"}free, spam\text{"}, \text{"}free, not\ spam\text{"}\}, 1) = 0$.