Notions of Probability

Madson Dias

@omadson

1 Machine Learning Notation

2 Probability

2.1 Sample space
2.2 Axioms of Probability
2.3 Conditional probability
2.4 Product rule
2.5 Total probability rule
2.6 Bayes Theorem
2.7 Independence

3 Random Variables and Distributions

Machine Learning Notation

Numbers and arrays

a	A scalar (integer or real)
A	A scalar constant
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix
$\boldsymbol{I}$	The $n \times n$ identity matrix
$\mathrm{diag}(\boldsymbol{a})$	A diagonal matrix with diagonal entries given by $\boldsymbol{a}$

Indexing

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1
a_{-i}	All elements of vector a except for element i
$A_{i,j}$	Element (i, j) of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{i,:}$	Row i of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{:,i}$	Column i of matrix $\boldsymbol{A}$

Linear Algebra

$\boldsymbol{A}^\top$	Transpose of matrix $\boldsymbol{A}$
$\boldsymbol{A} \odot \boldsymbol{B}$	Element-wise (Hadamard) product of $\boldsymbol{A}$ and $\boldsymbol{B}$
$\det(\boldsymbol{A})$	Determinant of $\boldsymbol{A}$
$\mathrm{Tr}(\boldsymbol{A})$	Trace of $\boldsymbol{A}$

Sets

$\mathcal{A}$ or $\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$[a, b]$	The real interval including a and b ($a \leq x \leq b$)
$(a, b]$	The real interval excluding a but including b ($a < x \leq b$)
$\mathcal{A} \setminus \mathcal{B}$	Set subtraction
$\mathcal{A} \cup \mathcal{B}$	Set union
$\mathcal{A} \cap \mathcal{B}$	Set intersection

Functions

$f : \mathcal{A} \to \mathcal{B}$	A function f with domain $\mathcal{A}$ and range $\mathcal{B}$
$f \circ g$	Composition of functions f and g
$f(\boldsymbol{x}; \boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$
$\ln x$ or $\log x$	Natural logarithm of $\boldsymbol{x}$
$\lVert \boldsymbol{x} \rVert_p$	L_p norm of $\boldsymbol{x}$
$\lVert \boldsymbol{x} \rVert$	L_2 norm of $\boldsymbol{x}$

Calculus

$f'(a)$ or $\frac{df}{dx}(a)$	Derivative of $f : \mathbb{R} \to \mathbb{R}$ at input point a
$\frac{\partial f}{\partial x_i}(\boldsymbol{a})$	Partial derivative of $f : \mathbb{R}^D \to \mathbb{R}$ with respect to x_i at input $\boldsymbol{a}$
$\nabla f(\boldsymbol{a})$	Gradient of $f : \mathbb{R}^D \to \mathbb{R}$ at input $\boldsymbol{a}$

Machine learning

$\boldsymbol{x}_n \in \mathbb{R}^D$	A D-dimensional input sample
$\boldsymbol{y}_n \in \mathbb{R}^S$	A S-dimensional output sample
$\hat{\boldsymbol{y}}_n \in \mathbb{R}^S$	An output (or label) predicted by a function f
$\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$	The training set (supervised learning)
$\mathcal{X} = \{\boldsymbol{x}_n\}_{n=1}^N$	The training set (unsupervised learning)

Probability

Quantitative description of a phenomenon whose outcome is **uncertain**.

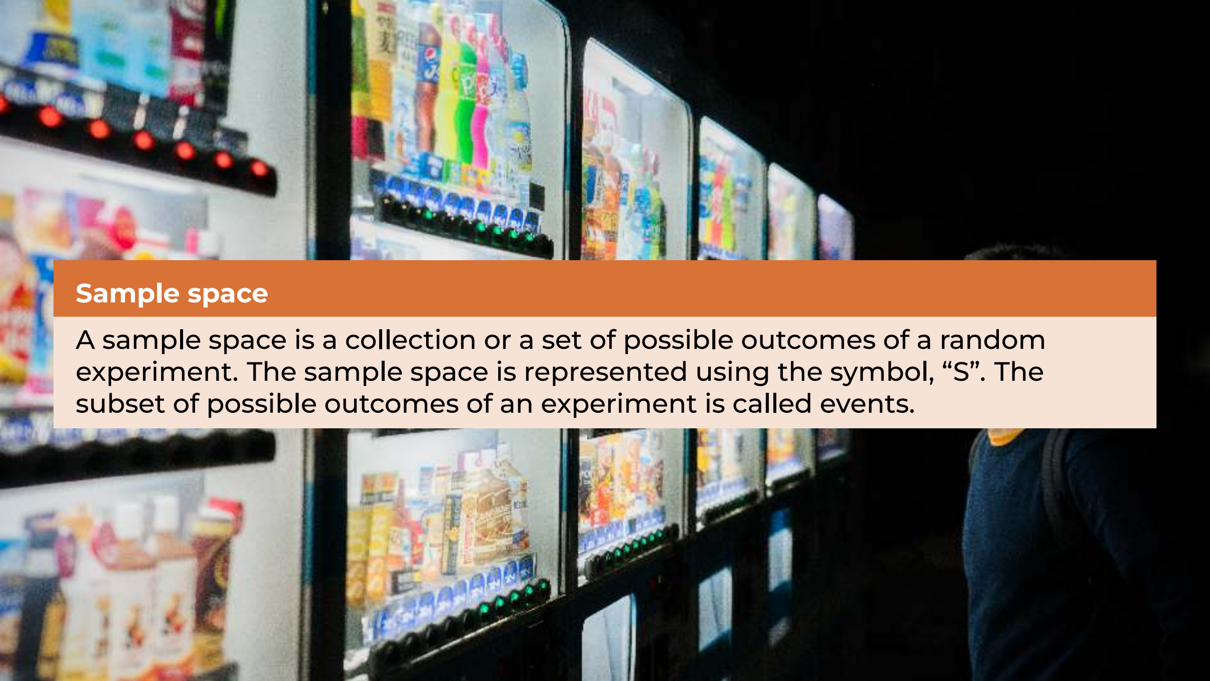
Types of phenomenon/experiments:

- **Deterministic**: it is a phenomenon that when repeated maintaining the same conditions, leads to the same result.
- **Random**: it is a phenomenon that, when repeated under the same conditions, it is impossible to predict its outcome.

How can I predict the result of a random event? I can’t!

Model to help predict the probability of some events.

1. Describe the possible results of the phenomenon
2. Describe our beliefs about the different possible results



Sample space

A sample space is a collection or a set of possible outcomes of a random experiment. The sample space is represented using the symbol, “S”. The subset of possible outcomes of an experiment is called events.

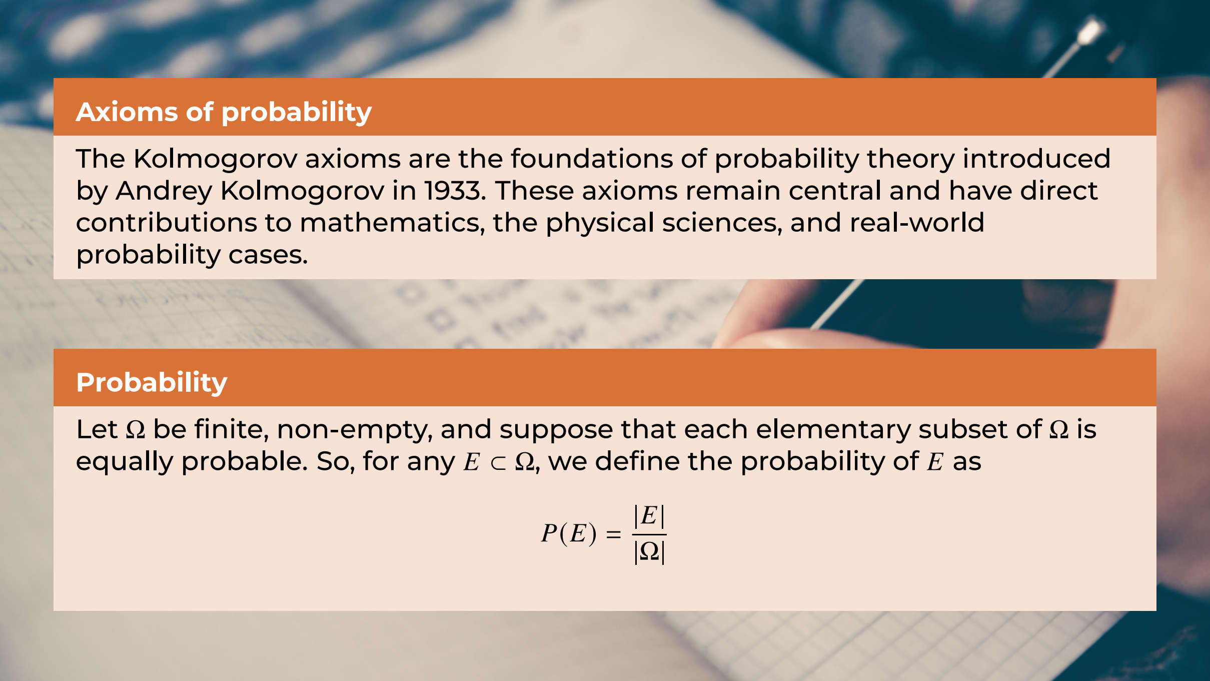
Set of possible outcomes of a experiment

The sample space, usually denoted by Ω or S, has elements that must be:

- mutually exclusive and
- collectively exhaustive.

Any subset E of the sample space Ω is known as an event

Two coin flipping example: {(head, tail), (tail, head), (head, head), (tail, tail)}



Axioms of probability

The Kolmogorov axioms are the foundations of probability theory introduced by Andrey Kolmogorov in 1933. These axioms remain central and have direct contributions to mathematics, the physical sciences, and real-world probability cases.

Probability

Let Ω be finite, non-empty, and suppose that each elementary subset of Ω is equally probable. So, for any $E \subset \Omega$, we define the probability of E as

$$P(E) = \frac{|E|}{|\Omega|}$$

Axiom 1 - Every probability is between 0 and 1 included, i.e:

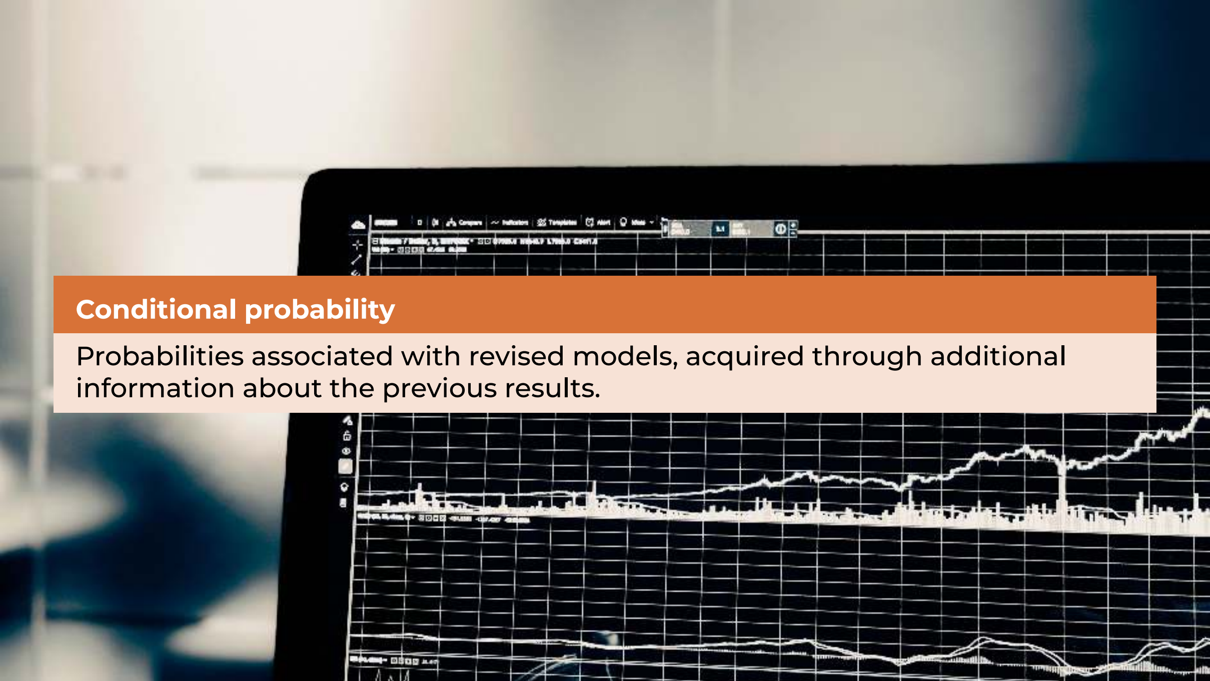
$$0 \leqslant P(E) \leqslant 1$$

Axiom 2 - The probability that at least one of the elementary events in the entire sample space Ω will occur is 1, i.e:

$$P(\Omega) = 1$$

Axiom 3 - For any sequence of mutually exclusive events $E_1, ..., E_N$, we have:

$$P\left(\bigcup_{n=1}^{N} E_n\right) = \sum_{n=1}^{N} P(E_n)$$



Conditional probability

Probabilities associated with revised models, acquired through additional information about the previous results.

Example #1: New clinical test for diabetes

A scientist is creating a new method for diagnosing diabetes

Experiment with 1000 patients (80 with diabetes)

The scientist then examined all patients using his method.

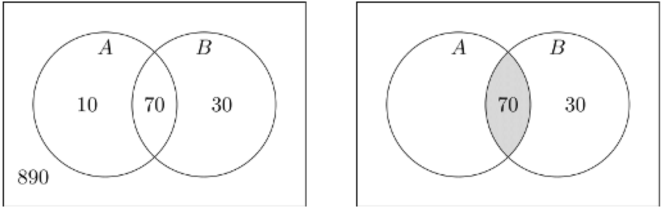
- 100 positive (and 900 negative)
- Only 70 actually had the disease.

How can I measure the effectiveness of this test?

Let’s consider the following events:

- A: the patient has diabetes;
- B: the test result is positive.

Probability of the occurrence of A, knowing that event B has already occurred, i.e., $P(A|B) = ?$



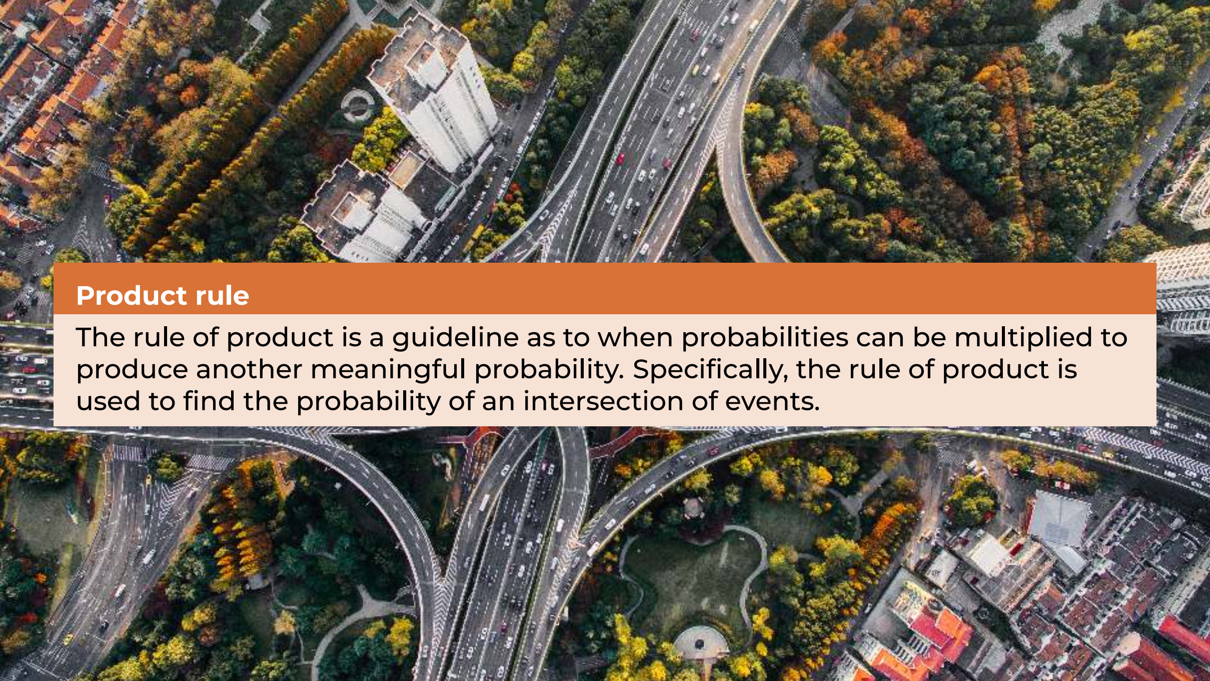
$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{70}{100} = 0.7 = 70\%$$

Conditional Probability

Let $A, B \in \Omega$ two events, the conditional probability of A given B is defined by

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

defined only in $P(B) > 0$.



Product rule

The rule of product is a guideline as to when probabilities can be multiplied to produce another meaningful probability. Specifically, the rule of product is used to find the probability of an intersection of events.

Example # 2: Improving estimates with more information

Same experiment of Example #1

450 from 1000 patients are male:

- 25 positive for the test with 20 positive for diabetes

How likely are you to have diabetes, being male and with a positive test?

To start we have to create an Event C: be male

What we would like to calculate is $P(A|B \cap C)$. That because it is expressed by

$$P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} \Leftrightarrow P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B|C) \cdot P(C)}.$$

Replacing

$$P(A|B \cap C) = \frac{\frac{20}{1000}}{\frac{25}{450} \cdot \frac{450}{1000}} \Leftrightarrow P(A|B \cap C) = \frac{4}{5} = 80\%.$$

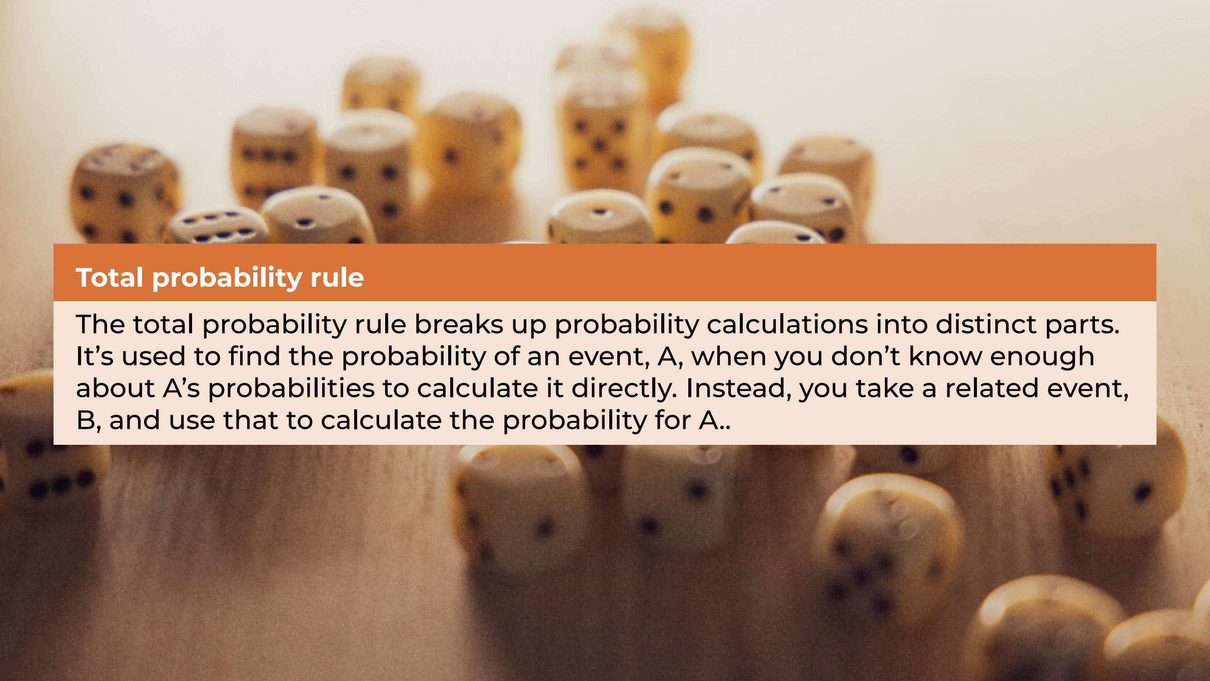
Changes in beliefs according to new evidence

Product rule

Suppose that $A_1, A_2, \ldots, A_N$ are arbitrary events. Then,

$$P(A_1 \cap A_2 \cap \cdots \cap A_N) = \prod_{n=1}^{N} P(A_n | A_1 \cap A_2 \cap \cdots \cap A_{n-1}),$$

$$P\left(\bigcap_{n=1}^{N} A_n\right) = \prod_{n=1}^{N} P\left(A_n \,\middle|\, \bigcap_{i=1}^{n-1} A_i\right).$$



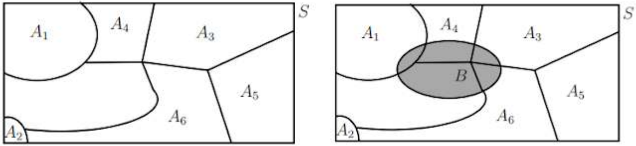
Total probability rule

The total probability rule breaks up probability calculations into distinct parts. It’s used to find the probability of an event, A, when you don’t know enough about A’s probabilities to calculate it directly. Instead, you take a related event, B, and use that to calculate the probability for A..

Consider a sample space divided into several non overlapping zones (left)

The sum of all these zones is equal to the sample space.

Now consider an event B (right).



B can be written as $B = (B \cap A_1) \cup \cdots \cup (B \cap A_N)$ and, with a little more math

$$P(B) = P(B|A_1) \cdot P(A_1) + \cdots + P(B|A_N) \cdot P(A_N).$$

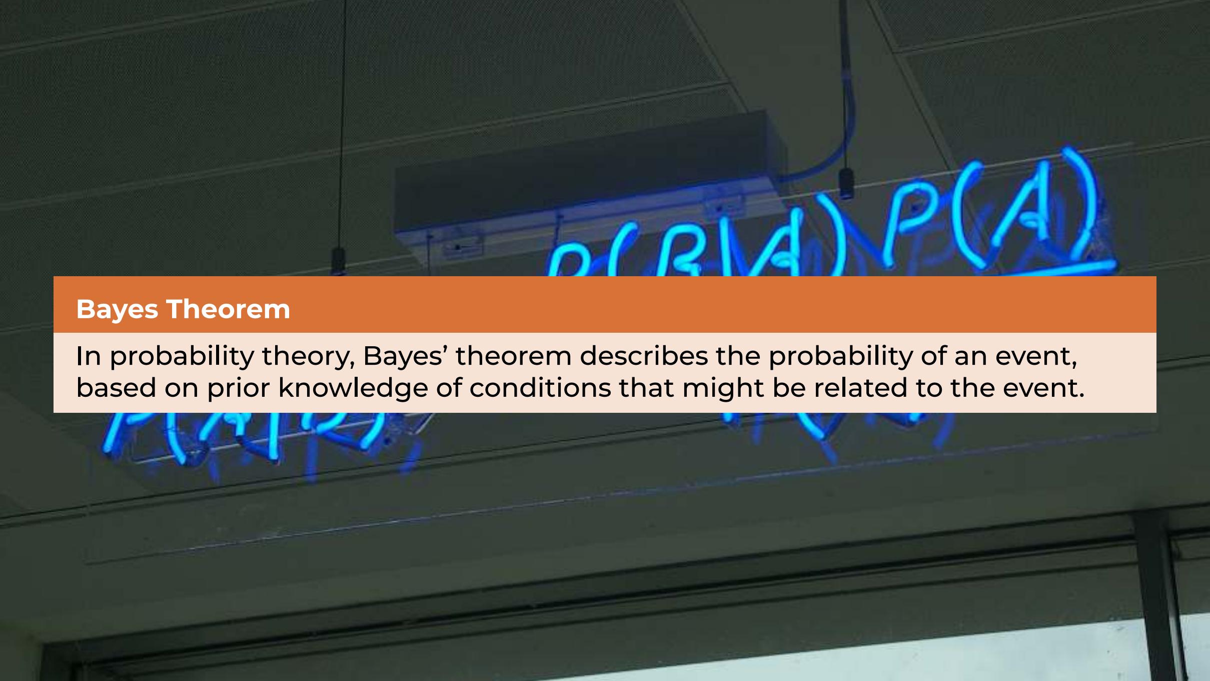
Total probability rule

Let B be any event, and let $A_1, A_2, \ldots, A_N$ be an arbitrary collection of events that satisfy the following properties:

1. $A_1 \cup \cdots \cup A_N = \Omega$
2. $A_i \cap A_j = \emptyset$

Then, the probability of event B can be calculated using the equation

$$P(B) = \sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)$$



Bayes Theorem

In probability theory, Bayes’ theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

Example #3: The two boxes

Box I: 2 white balls and 2 black balls

Box II: 1 white ball and 3 black balls

Experiment:

1. Selecting a ball from box I and transferring it to the second
2. Then a ball is selected in box II

What is the probability that the ball selected in box I was white, given that the selected ball of box II is white?

The first thing we must do is to define the events:

A: take out a white ball in box I

B: take out a white ball in box II

Our question now is $P(A|B) = ?$

Compute the conditional probability $P(A|B)$

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

Replace the term $P(A \cap B)$ with its corresponding $P(B \cap A)$

$$P(A|B) = \frac{P(B \cap A)}{P(B)} \Leftrightarrow P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

It is easy to see that $P(B|A) = \frac{2}{5}$ and $P(A) = \frac{2}{4}$.

Note that $P(B)$ changes according to the color of the ball taken in box I.

We only have two options: A or $\bar{A}$.

$$\begin{aligned} P(B) &= P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A}) \\ &= \frac{2}{5} \cdot \frac{1}{2} + \frac{1}{5} \cdot \frac{1}{2} \Leftrightarrow P(B) = \frac{3}{10}. \end{aligned}$$

Therefore

$$\begin{aligned} P(A|B) &= \frac{\frac{2}{5} \cdot \frac{2}{4}}{\frac{3}{10}}, \\ &= \frac{2}{3}. \end{aligned}$$

In the first part of the example we manipulate conditional probabilities until the simplest form of *Bayes Rule*

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

This rule describes the probability of an event, based on knowledge *a priori* that may be related to the event.

Bayes Theorem

Let $\{A_1, A_2, \ldots, A_N\}$ be a partition of the sample space Ω in events of positive probability. If B is an event with $P(B) > 0$, then, for every $n = 1, \ldots, N$,

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{n=1}^{N} P(B|A_n)P(A_n)}.$$

Independence

Two events A and B are independent if and only if we have:

$$P(A \cap B) = P(A) \cdot P(B).$$

Random Variables and Distributions

A variable whose possible values are the result of a random phenomenon.

Maps the results of an unpredictable process across numerical quantities.

Classification: Continuous/Discrete; Univariate / Multivariate.

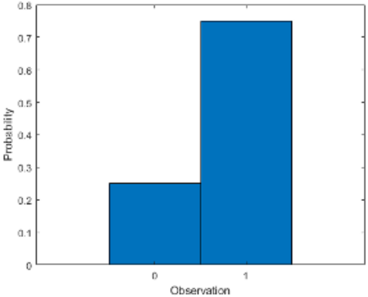
Formally: A random variable, often noted X, is a function that maps every element in a sample space to a real line.

Discrete probabilities/Random Variables

X: Result of a coin toss: $X \sim B(p)$

$$P(X = 1) = p \text{ or } P(X = 0) = 1 - p$$

- Bernoulli distribution:



$$f(k;p) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}$$

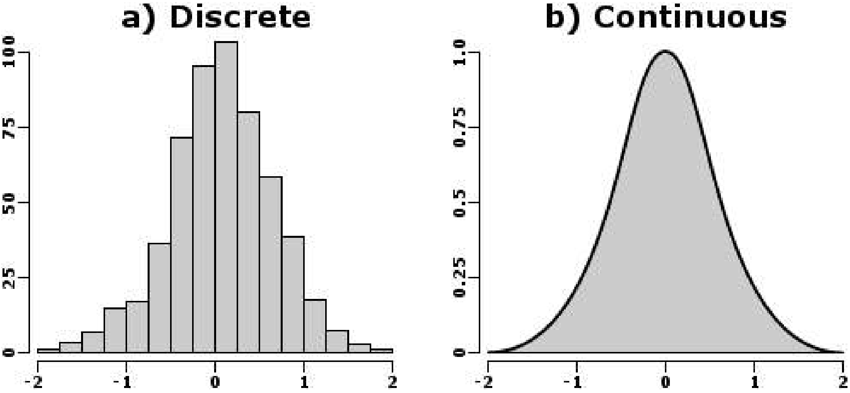
Continuous probabilities/Random Variables

- How likely is it to hit the center of the target?

$$P(X = x \cap Y = y) = P(X = x, Y = y) = 0$$

Discrete: Probability mass function (PMF)

Continuous: Probability density function (PDF)



Expectation

The expected value (or mean) of X, where X is a discrete random variable, is a weighted average of the possible values that X can take, each value being weighted according to the probability of that event occurring, usually written as $\mathrm{E}[X]$ or μ.

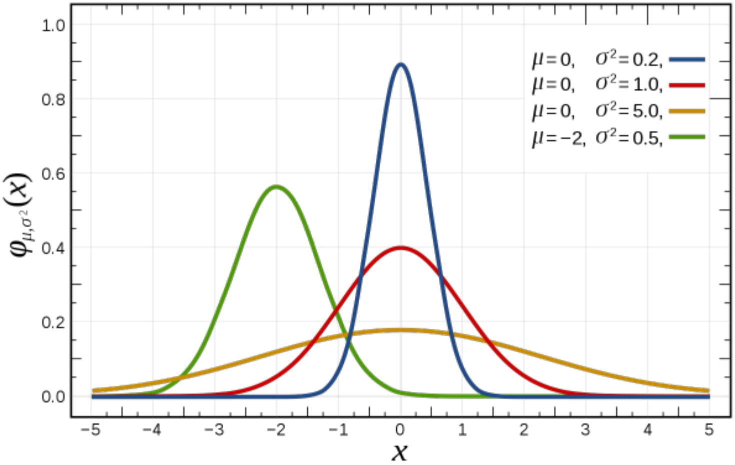
$$\mathrm{E}[X] = \sum_{x \in \Omega} x \cdot P(X = x) \rightarrow \mathrm{E}[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

Variance

Variance is the expectation of the squared deviation of a random variable from its mean

$$\mathrm{Var}[X] = \sum_{x \in \Omega} P(X = x)(x - \mu) \rightarrow \mathrm{Var}[X] = \mathrm{E}[(X - \mu)^2],$$

Example of distributions with different expectation and variance[1]



[1] σ^2 is the variance

omadson.codes