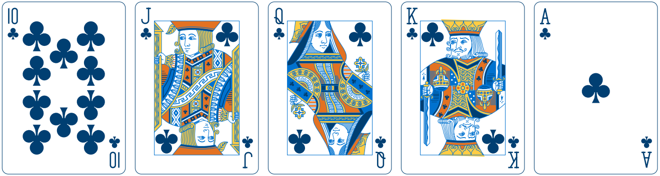
2P7: Probability & Statistics

Probability Fundamentals

Thierry Savin

Lent 2024



the *royal flush*, the best possible hand in poker, has a probability 0.000154%

Probability: mathematics of uncertain events

Statistics: science of collecting and analysing data

- Probability is logically self-contained
  - Few rules
  - Answers all follow logically from the rules
  - Computations can be tricky (but rarely messy)
  - Example: a fair coin is tossed 100 times, what is the probability of 60 or more heads?
- Statistics apply probability to draw conclusions from data
  - Computations can be messy (and tricky)
  - Example: an unknown coin is tossed 100 times and lands 60 heads, what can we conclude about its fairness?

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Central questions of the student:

- Why do we need this?
  - Make inference about uncertain events
  - Test the strength of statistical evidence
  - Form the basis of many other theories
- How is it possible to say something about uncertain events? How can we measure uncertainty?
- How do I get a good mark?

Examples of applications: failure analysis, design, risk assessment, reliability theory, environmental regulations, inventory theory, computing and simulation, mathematical finance, queueing theory, disease spread, clinical trials, quantum physics, telecommunication, traffic engineering, fitting and machine learning, neural dynamics, statistical mechanics, . . .

Probability and statistics: important tools in all kinds of Engineering!

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Seven lectures (weeks 1-4) to cover the following:

1. Probability Fundamentals
2. Discrete Probability Distributions
3. Continuous Random Variables
4. Manipulating and Combining Distributions
5. Decision, Estimation and Hypothesis Testing

Introduction (*what we're doing now...*)

Foundations of Probability

Conditional Probability

Discrete random variables

Expectation and Entropy

UNIVERSITY OF CAMBRIDGE
Department of Engineering

- In classical frequentist statistics, the probability of an event is defined as "its long-run frequency in a repeatable experiment".

  Example: "the probability of rolling a 6 with a fair dice is $\frac{1}{6}$" because this is the relative frequency of this event as the number of experiments tends to infinity.

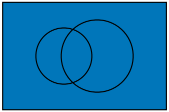
- However, some notions of chance don't lend themselves to a frequentist approach, and an interpretation of probability with a (subjective) degree of belief is possible; this is known as the Bayesian interpretation.

  Example: "there is a 50% chance that the arctic polar ice cap will have melted by the year 2100", it is not possible to define a repeatable experiment.

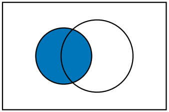
Both approaches can be treated using the same probability theory.

UNIVERSITY OF CAMBRIDGE
Department of Engineering

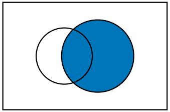
- A set Ω is a collection of elements.
- Element: We write $\omega \in \Omega$ to mean the element ω is in the set Ω.
- Subset: We say the set $\mathcal{A}$ is a subset of Ω if all of its elements are in Ω. We write this as $\mathcal{A} \subset \Omega$ ($\mathcal{A} \subseteq \Omega$ if $\mathcal{A} = \Omega$ possible).
- Complement: The complement of $\mathcal{A}$ in Ω is the set of elements of Ω that are not in $\mathcal{A}$. We write this as $\mathcal{A}^{\complement}$.
- Union: The union of $\mathcal{A}$ and $\mathcal{B}$ is the set of all elements in $\mathcal{A}$ <u>or</u> $\mathcal{B}$ <u>or</u> both. We write this as $\mathcal{A} \cup \mathcal{B}$.
- Intersection: The intersection of $\mathcal{A}$ and $\mathcal{B}$ is the set of all elements in both $\mathcal{A}$ <u>and</u> $\mathcal{B}$. We write this as $\mathcal{A} \cap \mathcal{B}$.
- Empty set: The empty set is the set with no elements. We denote it $\varnothing$. If $\mathcal{A} \cap \mathcal{B} = \varnothing$, $\mathcal{A}$ and $\mathcal{B}$ are said to be *disjoint*.



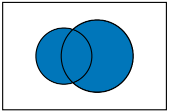
Ω



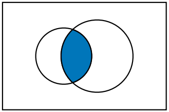
$\mathcal{A}$



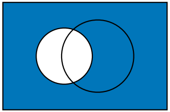
$\mathcal{B}$



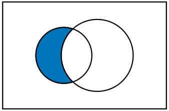
$\mathcal{A} \cup \mathcal{B}$



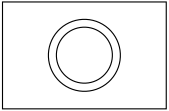
$\mathcal{A} \cap \mathcal{B}$



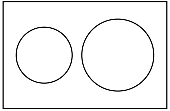
$\mathcal{A}^{\mathsf{c}}$



$\mathcal{A} \cap \mathcal{B}^{\mathsf{c}}$



$\mathcal{A} \subset \mathcal{B}$



$\mathcal{A} \cap \mathcal{B} = \varnothing$

Examples: prove De Morgan's laws

$$(\mathcal{A} \cup \mathcal{B})^{\mathsf{c}} = \mathcal{A}^{\mathsf{c}} \cap \mathcal{B}^{\mathsf{c}}$$
$$(\mathcal{A} \cap \mathcal{B})^{\mathsf{c}} = \mathcal{A}^{\mathsf{c}} \cup \mathcal{B}^{\mathsf{c}}$$

UNIVERSITY OF CAMBRIDGE
Department of Engineering

- Experiment: a repeatable procedure with well-defined[1] possible outcomes.
- Sample space: the *set* of all possible outcomes; noted Ω.
- Event: a *subset* of the sample space, $\mathcal{A} \subseteq \Omega$.

Examples:

- Experiment 1: toss a fair coin, report if it lands heads or tails.
  - $\Omega = \{\text{H,T}\}$.
  - $\mathcal{A} =$ "heads" $= \{\text{H}\}$ (single *element*).
- Experiment 2: toss a fair coin 3 times, list the results.
  - $\Omega = \{\text{HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}\}$
  - $\mathcal{A} =$ "exactly 2 heads" $= \{\text{HHT,HTH,THH}\}$.

[1] each outcome is unique, and different outcomes are *mutually exclusive*

UNIVERSITY OF CAMBRIDGE
Department of Engineering

The probability $\mathbb{P}$ is a *measure* that verifies the following:

- The probability of an event is a non-negative real number[2]

$$\mathbb{P}[\mathcal{A}] \in \mathbb{R} \quad \text{and} \quad \mathbb{P}[\mathcal{A}] \geq 0, \quad \text{for all } \mathcal{A} \subseteq \Omega$$

- The sample space (also called "certain event") has unit probability

$$\mathbb{P}[\Omega] = 1$$

- Additivity for *incompatible* events (i.e. disjoint sets):

$$\mathbb{P}[\mathcal{A} \cup \mathcal{B}] = \mathbb{P}[\mathcal{A}] + \mathbb{P}[\mathcal{B}] \quad \text{if } \mathcal{A} \cap \mathcal{B} = \varnothing$$

Note: it is OK to write $\mathbb{P}[\mathcal{A}]$ as $P[A]$ in your handwritten notes/exam.

[2] $\mathbb{R}$ denotes the set of real numbers.

UNIVERSITY OF CAMBRIDGE
Department of Engineering

These three axioms are sufficient to derive[3] the following:

- Monotonicity
$$\text{if} \quad \mathcal{A} \subseteq \mathcal{B} \quad \text{then} \quad \mathbb{P}[\mathcal{A}] \leq \mathbb{P}[\mathcal{B}]$$
- Probability of the empty set
$$\mathbb{P}[\varnothing] = 0$$
- Complement rule
$$\mathbb{P}[\mathcal{A}^{\complement}] = 1 - \mathbb{P}[\mathcal{A}]$$
- Numeric bound
$$0 \leq \mathbb{P}[\mathcal{A}] \leq 1, \quad \text{for all } \mathcal{A} \subseteq \Omega$$
- Addition law
$$\mathbb{P}[\mathcal{A} \cup \mathcal{B}] = \mathbb{P}[\mathcal{A}] + \mathbb{P}[\mathcal{B}] - \mathbb{P}[\mathcal{A} \cap \mathcal{B}]$$
- Sum rule
$$\mathbb{P}[\mathcal{A} \cap \mathcal{B}] + \mathbb{P}[\mathcal{A} \cap \mathcal{B}^{\complement}] = \mathbb{P}[\mathcal{A}]$$

[3] proofs in the examples paper

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Another consequence of the additivity axiom is

$$\mathbb{P}[\mathcal{A}] = \sum_{\omega \in \mathcal{A}} \mathbb{P}[\omega]$$

where $\{\omega \in \Omega\}$ are the mutually-disjoint *individual outcomes* (sometimes called "atomic events") of the experiment.

- These calculations are often done with *combinatorics*;
- They inform on the "physical" origin of the probability of an event;
- We won't do much of these here.

Conditional probability answers the question "how does the probability of an event change if we have extra information?".

Example: Toss a fair coin 3 times.

- What is the probability of 3 heads?

  $\Omega = \{\text{HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}\}$, with all outcomes equally likely, so with $\mathcal{A} = \{\text{HHH}\}$
  $$\mathbb{P}[\mathcal{A}] = 1/8$$

- Given the 1st toss is heads, what is the probability of 3 heads?

  Reduced sample space to $\mathcal{B} = \{\text{HHH,HHT,HTH,HTT}\}$, with all outcomes equally likely, so
  $$\mathbb{P}[\text{“ } \mathcal{A} \text{ occurs given } \mathcal{B} \text{ occurred ”}] = 1/4$$

The conditional probability of an event $\mathcal{A}$ knowing that an event $\mathcal{B}$ occurred is written:
$$\mathbb{P}[\mathcal{A}|\mathcal{B}] \quad (\text{“ } \mathcal{A} \text{ given } \mathcal{B} \text{ ”})$$

UNIVERSITY OF CAMBRIDGE
Department of Engineering

The formal definition of conditional probability[4] reads:

$$\mathbb{P}[\mathcal{A}|\mathcal{B}] = \frac{\mathbb{P}[\mathcal{A} \cap \mathcal{B}]}{\mathbb{P}[\mathcal{B}]}, \quad \text{provided } \mathbb{P}[\mathcal{B}] \neq 0$$

Why? Frequency interpretation:

- Repeat experiment N times keeping track of events $\mathcal{A}$ and $\mathcal{B}$.
- $N_{\mathcal{B}}$ number of times $\mathcal{B}$ is realised:
$$\mathbb{P}[\mathcal{B}] \overset{N\to\infty}{\approx} \frac{N_{\mathcal{B}}}{N}$$
- $N_{\mathcal{A}\cap\mathcal{B}}$ number of times both $\mathcal{A}$ and $\mathcal{B}$ occur:
$$\mathbb{P}[\mathcal{A} \cap \mathcal{B}] \overset{N\to\infty}{\approx} \frac{N_{\mathcal{A}\cap\mathcal{B}}}{N}$$
- Conditional probability:
$$\mathbb{P}[\mathcal{A}|\mathcal{B}] \overset{N_{\mathcal{B}}\to\infty}{\approx} \frac{N_{\mathcal{A}\cap\mathcal{B}}}{N_{\mathcal{B}}}$$
" proportion of $\mathcal{A}$ among the occurrences of $\mathcal{B}$ "

[4] one can show that the conditional probability follows the 3 axioms

UNIVERSITY OF CAMBRIDGE
Department of Engineering

- Product rule

$$\mathbb{P}[\mathcal{A} \cap \mathcal{B}] = \mathbb{P}[\mathcal{A}|\mathcal{B}]\,\mathbb{P}[\mathcal{B}]$$

Proof: from the definition of conditional probability $\square$

- Law of Total Probability

$$\mathbb{P}[\mathcal{A}] = \mathbb{P}[\mathcal{A}|\mathcal{B}]\,\mathbb{P}[\mathcal{B}] + \mathbb{P}[\mathcal{A}|\mathcal{B}^{\complement}]\,\mathbb{P}[\mathcal{B}^{\complement}]$$

Proof: from the sum rule $\square$

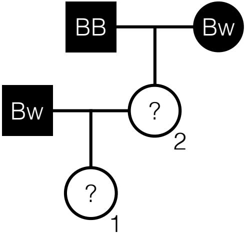
- More generally, with $\{\mathcal{B}_k : k = 1, 2, 3, \ldots n\}$ a set of *pairwise incompatible* events with[5] $\bigcup_{k=1}^{n} \mathcal{B}_k = \Omega$,

$$\mathbb{P}[\mathcal{A}] = \sum_{k=1}^{n} \mathbb{P}[\mathcal{A}|\mathcal{B}_k]\,\mathbb{P}[\mathcal{B}_k]$$

[5] We denote $\bigcup_{k=1}^{n} \mathcal{B}_k = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_n$.

Conditional Probability

Example: Pedigree



- Female (circle) and male (square) cats transmit each one colour gene to offspring
- Black (B) dominant, white (w) recessive
- What is the probability that cat 1 is heterozygous (carries both genes) if it is black?

$\mathcal{B}_i$: "cat i is black"
$\mathcal{H}_j$: "cat j is heterozygous"

What is $\mathbb{P}[\mathcal{H}_1|\mathcal{B}_1]$?

$$\begin{aligned}\mathbb{P}[\mathcal{B}_1] &= \mathbb{P}[\mathcal{B}_1|\mathcal{H}_2] \times \mathbb{P}[\mathcal{H}_2] + \mathbb{P}[\mathcal{B}_1|\mathcal{H}_2^\complement] \times \mathbb{P}[\mathcal{H}_2^\complement] \\ &= \frac{3}{4} \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{7}{8}\end{aligned}$$

$$\begin{aligned}\mathbb{P}[\mathcal{H}_1] &= \mathbb{P}[\mathcal{H}_1|\mathcal{H}_2] \times \mathbb{P}[\mathcal{H}_2] + \mathbb{P}[\mathcal{H}_1|\mathcal{H}_2^\complement] \times \mathbb{P}[\mathcal{H}_2^\complement] \\ &= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}\end{aligned}$$

$$\mathbb{P}[\mathcal{H}_1|\mathcal{B}_1] = \frac{\mathbb{P}[\mathcal{H}_1 \cap \mathcal{B}_1]}{\mathbb{P}[\mathcal{B}_1]} = \frac{\mathbb{P}[\mathcal{H}_1]}{\mathbb{P}[\mathcal{B}_1]} = \frac{4}{7} \qquad \square$$

UNIVERSITY OF CAMBRIDGE
Department of Engineering

- Bayes' rule

$$\mathbb{P}[\mathcal{B}|\mathcal{A}] = \frac{\mathbb{P}[\mathcal{A}|\mathcal{B}]\,\mathbb{P}[\mathcal{B}]}{\mathbb{P}[\mathcal{A}]}$$

- Bayes'rule tells us how to invert conditional probabilities, i.e. to find $\mathbb{P}[\mathcal{B}|\mathcal{A}]$ from $\mathbb{P}[\mathcal{A}|\mathcal{B}]$.
- Proof via the *product rule*

$$\begin{aligned}\mathbb{P}[\mathcal{A}\cap\mathcal{B}] &= \mathbb{P}[\mathcal{A}|\mathcal{B}]\,\mathbb{P}[\mathcal{B}]\\ &= \mathbb{P}[\mathcal{B}\cap\mathcal{A}]\\ &= \mathbb{P}[\mathcal{B}|\mathcal{A}]\,\mathbb{P}[\mathcal{A}] \quad \square\end{aligned}$$

- Often used with the *law of total probability*

$$\mathbb{P}[\mathcal{B}|\mathcal{A}] = \frac{\mathbb{P}[\mathcal{A}|\mathcal{B}]\,\mathbb{P}[\mathcal{B}]}{\mathbb{P}[\mathcal{A}|\mathcal{B}]\,\mathbb{P}[\mathcal{B}] + \mathbb{P}[\mathcal{A}|\mathcal{B}^{\complement}]\,\mathbb{P}[\mathcal{B}^{\complement}]}$$

UNIVERSITY OF CAMBRIDGE
Department of Engineering

The rapid antigen test for covid has a sensitivity of 78% (true positive rate), and a specificity of 97% (true negative rate). The population sees 5% incidence of covid.

What is the probability that a person testing positive has covid?

$\mathcal{T}$: The test is positive

$\mathcal{C}$: The person has covid

What is $\mathbb{P}[\mathcal{C}|\mathcal{T}]$?

The data tell us:

$$\mathbb{P}[\mathcal{T}|\mathcal{C}] = 0.78 \qquad \mathbb{P}[\mathcal{T}|\mathcal{C}^{\complement}] = 0.03$$
$$\mathbb{P}[\mathcal{T}^{\complement}|\mathcal{C}] = 0.22 \qquad \mathbb{P}[\mathcal{T}^{\complement}|\mathcal{C}^{\complement}] = 0.97$$
$$\mathbb{P}[\mathcal{C}] = 0.05 \qquad \mathbb{P}[\mathcal{C}^{\complement}] = 0.95$$

Now we write:

$$\mathbb{P}[\mathcal{C}|\mathcal{T}] = \frac{\mathbb{P}[\mathcal{T}|\mathcal{C}]\,\mathbb{P}[\mathcal{C}]}{\mathbb{P}[\mathcal{T}]} = \frac{\mathbb{P}[\mathcal{T}|\mathcal{C}]\,\mathbb{P}[\mathcal{C}]}{\mathbb{P}[\mathcal{T}|\mathcal{C}]\,\mathbb{P}[\mathcal{C}] + \mathbb{P}[\mathcal{T}|\mathcal{C}^{\complement}]\,\mathbb{P}[\mathcal{C}^{\complement}]}$$
$$= \frac{0.78 \times 0.05}{0.78 \times 0.05 + 0.03 \times 0.95} = 58\% \qquad \square$$

UNIVERSITY OF CAMBRIDGE
Department of Engineering

- Two events are *independent* if the knowledge that one occurred does not change the probability that the other occurs. Independence is fundamental, as we shall see later.
- In mathematical terms, that means that $\mathcal{A}$ and $\mathcal{B}$ are independent if $\mathbb{P}[\mathcal{A}|\mathcal{B}] = \mathbb{P}[\mathcal{A}]$. Using the product rule, we arrive at the ...
- Formal definition of independence:

$$\mathcal{A} \text{ and } \mathcal{B} \text{ independent} \quad \Leftrightarrow \quad \mathbb{P}[\mathcal{A} \cap \mathcal{B}] = \mathbb{P}[\mathcal{A}] \times \mathbb{P}[\mathcal{B}]$$

- Frequency interpretation:

$$\frac{N_{\mathcal{A}}}{N} \overset{N\to\infty}{\approx} \mathbb{P}[\mathcal{A}] = \frac{\mathbb{P}[\mathcal{A} \cap \mathcal{B}]}{\mathbb{P}[\mathcal{B}]} \overset{N\to\infty}{\approx} \frac{N_{\mathcal{A}\cap\mathcal{B}}}{N} \times \frac{N}{N_{\mathcal{B}}} = \frac{N_{\mathcal{A}\cap\mathcal{B}}}{N_{\mathcal{B}}}$$

Knowing when $\mathcal{B}$ occurs is irrelevant to the probability of $\mathcal{A}$.

- **Formal definition**:
  - A *discrete* random variable X is a function of the outcomes of a random experiment, $\mathrm{X} : \Omega \to \mathbb{X}$, that takes a *discrete* set of scalar values forming $\mathbb{X}$ (typically $\mathbb{X}$ is a subset of $\mathbb{R}$, the set or real numbers).
  - $\mathbb{X}$ is called the *support* (or sometimes *alphabet*) of X.
- **Example: game with 2 dice**
  - Roll a dice twice and record the outcomes (i, j) with i and j the result of the 1st and 2nd roll, respectively.
  - $\Omega = \{(1,1),(1,2),(1,3),\ldots,(6,6)\}$ (36 outcomes)
  - We define the random variable $\mathrm{X}(i,j) = \max\{i,j\}$.
  - The event $\mathrm{X} = 3$ is $\{(1,3),(2,3),(3,3),(3,2),(3,1)\}$ and $\mathbb{P}[\mathrm{X} = 3] = \frac{5}{36}$.
  - We can build the table:

x	1	2	3	4	5	6	7 ...
$\mathbb{P}[\mathrm{X} = x]$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	0 ...

- The probability mass function (PMF) of a *discrete* random variable X is the function

$$P_{\mathrm{X}} : \mathbb{X} \to [0,1] \quad \text{with} \quad P_{\mathrm{X}}(x) = \mathbb{P}[\mathrm{X} = x]$$

- About the notations:
  - X designates the random variable. A random variable is often denoted by capital roman letters, such as X, Y, Z, T ...
  - x is the independent variable (arbitrary input) and I could use whatever symbol for it: $P_{\mathrm{X}}(a) = \mathbb{P}[\mathrm{X} = a]$ does not change the definition of P_{X} .
  - the name of the function is P_{X}, with the "subscript X" to remind us it is associated with the random variable X.
- Finding the PMF of a random variable comes from how the random variable originates from the outcomes.

Note: it is OK to write $P_{\mathrm{X}}(x)$ as $P_X(x)$ in your handwritten notes/exam.

UNIVERSITY OF CAMBRIDGE
Department of Engineering

- Note that the events "$\mathrm{X} = a$" and "$\mathrm{X} = b$" are disjoint if $a \neq b$. Hence $\sum_{x \in \mathbb{X}} P_{\mathrm{X}}(x) = 1$.
- For $x \notin \mathbb{X}$, we can set $P_{\mathrm{X}}(x) = 0$ (if x out of range, "$\mathrm{X} = x$" $= \varnothing$ the empty event)
- It will be useful to introduce the cumulative distribution function:

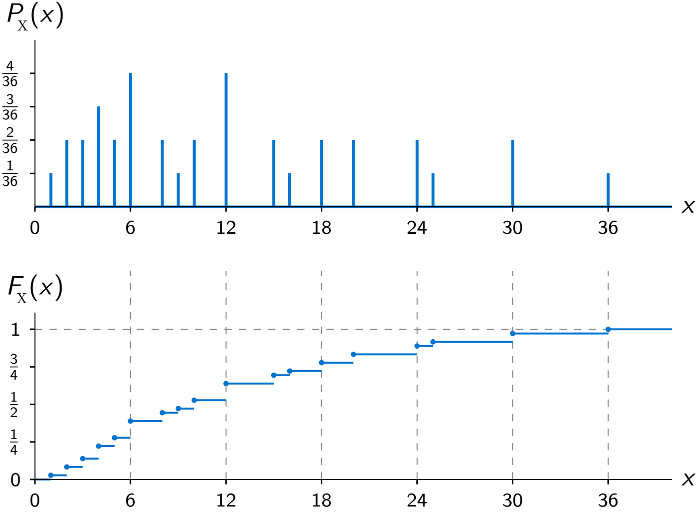
$$F_{\mathrm{X}}(x) = \mathbb{P}[\mathrm{X} \leq x]$$

It is obtained by adding up the probabilities $P_{\mathrm{X}}(\xi)$ as ξ runs from $-\infty$ to x, $F_{\mathrm{X}}(x) = \sum_{\xi \leq x} P_{\mathrm{X}}(\xi)$. It has the following properties:

  - F_{X} is non-decreasing: $F_{\mathrm{X}}(a) \leq F_{\mathrm{X}}(b)$ if $a \leq b$
  - $\lim_{x \to -\infty} F_{\mathrm{X}}(x) = 0$ and $\lim_{x \to \infty} F_{\mathrm{X}}(x) = 1$
  - $\mathbb{P}[a < \mathrm{X} \leq b] = F_{\mathrm{X}}(b) - F_{\mathrm{X}}(a)$
    using $\{\mathrm{X} \leq b\} = \{\mathrm{X} \leq a\} \cup \{a < \mathrm{X} \leq b\}$ $\square$

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Game with 2 dice, and $\mathrm{X}(i,j) = i \times j$



For two random variables $\mathrm{X} : \Omega \to \mathbb{X}$ and $\mathrm{Y} : \Omega \to \mathbb{Y}$ defined on the *same* sample space, we can introduce the joint probability mass function:

$$P_{\mathrm{XY}} : \mathbb{X} \times \mathbb{Y} \to [0,1] \quad \text{with} \quad P_{\mathrm{XY}}(x,y) = \mathbb{P}[\mathrm{X} = x \cap \mathrm{Y} = y]$$

And we inherit the following ideas from our previous discussion:

- Conditional probability
$$P_{\mathrm{X|Y}} : \mathbb{X} \times \mathbb{Y} \to [0,1] \quad \text{with} \quad P_{\mathrm{X|Y}}(x|y) = \frac{P_{\mathrm{XY}}(x,y)}{P_{\mathrm{Y}}(y)}$$
- Law of total probability, also known as marginalisation
$$P_{\mathrm{X}}(x) = \sum_{y \in \mathbb{Y}} P_{\mathrm{XY}}(x,y)$$
- Bayes' rule
$$P_{\mathrm{Y|X}}(y|x) = \frac{P_{\mathrm{X|Y}}(x|y)P_{\mathrm{Y}}(y)}{P_{\mathrm{X}}(x)} = \frac{P_{\mathrm{X|Y}}(x|y)P_{\mathrm{Y}}(y)}{\sum_{\xi \in \mathbb{Y}} P_{\mathrm{X|Y}}(x|\xi)P_{\mathrm{Y}}(\xi)}$$

Note: OK to write $P_{\mathrm{XY}}(x,y)$ as $P_{xy}(x,y)$ and $P_{\mathrm{X|Y}}(x|y)$ as $P_{x|y}(x|y)$.

- Two random variables X and Y are independent iff *all the events* corresponding to values of X are independent of *all the events* corresponding to values of Y:

$$\begin{aligned} P_{\mathrm{XY}}(x,y) &= \mathbb{P}[\mathrm{X}=x \cap \mathrm{Y}=y] \\ &= \mathbb{P}[\mathrm{X}=x] \times \mathbb{P}[\mathrm{Y}=y] \\ P_{\mathrm{XY}}(x,y) &= P_{\mathrm{X}}(x) P_{\mathrm{Y}}(y) \quad \text{for all } x,y \in \mathbb{X} \times \mathbb{Y} \end{aligned}$$

- For more than two random variables, we can also define a joint multivariate probability mass function $P_{\mathrm{X_1X_2\ldots X_n}}(x_1, x_2, \ldots, x_n)$. The variables are *mutually independent* iff

$$P_{\mathrm{X_1X_2\ldots X_n}}(x_1, x_2, \ldots, x_n) = P_{\mathrm{X_1}}(x_1) P_{\mathrm{X_2}}(x_2) \ldots P_{\mathrm{X_n}}(x_n)$$

for all $x_1, x_2, \ldots, x_n \in \mathbb{X}_1 \times \mathbb{X}_2 \times \ldots \mathbb{X}_n$.
This is different from *pairwise independence*, which only requires

$$P_{\mathrm{X_iX_j}}(x_i, x_j) = P_{\mathrm{X_i}}(x_i) P_{\mathrm{X_j}}(x_j) \quad \text{for all } i,j \text{ and all } x_i, x_j \in \mathbb{X}_i \times \mathbb{X}_j$$

UNIVERSITY OF CAMBRIDGE
Department of Engineering

The XOR gate

- Consider two independent binary random variables X and Y with $P_{\mathrm{X}}(x) = \frac{1}{2}$ if $x = 0$ or $x = 1$, and $P_{\mathrm{X}}(x) = 0$ otherwise. X and Y are *identically distributed*, so $P_{\mathrm{X}} = P_{\mathrm{Y}}$.
- A third random variable Z is obtained by $\mathrm{Z} = \mathrm{X}$ XOR Y (a XOR $b = 0$ when $a = b$, 1 otherwise).
- Show that $\mathrm{X}, \mathrm{Y}, \mathrm{Z}$ are *pairwise*, but <u>not</u> *mutually*, independent.

The joint distribution of $\mathrm{X}, \mathrm{Y}, \mathrm{Z}$:

$$
\begin{aligned}
P_{\mathrm{XYZ}}(0,0,0) &= \tfrac{1}{4}\\
P_{\mathrm{XYZ}}(0,0,1) &= 0\\
P_{\mathrm{XYZ}}(0,1,0) &= 0\\
P_{\mathrm{XYZ}}(0,1,1) &= \tfrac{1}{4}\\
P_{\mathrm{XYZ}}(1,0,0) &= 0\\
P_{\mathrm{XYZ}}(1,0,1) &= \tfrac{1}{4}\\
P_{\mathrm{XYZ}}(1,1,0) &= \tfrac{1}{4}\\
P_{\mathrm{XYZ}}(1,1,1) &= 0
\end{aligned}
$$

By marginalisation over Y:

$$
\begin{aligned}
P_{\mathrm{XZ}}(0,0) &= P_{\mathrm{XYZ}}(0,0,0) + P_{\mathrm{XYZ}}(0,1,0) = \tfrac{1}{4}\\
P_{\mathrm{XZ}}(0,1) &= P_{\mathrm{XYZ}}(0,0,1) + P_{\mathrm{XYZ}}(0,1,1) = \tfrac{1}{4}\\
P_{\mathrm{XZ}}(1,0) &= P_{\mathrm{XYZ}}(1,0,0) + P_{\mathrm{XYZ}}(1,1,0) = \tfrac{1}{4}\\
P_{\mathrm{XZ}}(1,1) &= P_{\mathrm{XYZ}}(1,0,1) + P_{\mathrm{XYZ}}(1,1,1) = \tfrac{1}{4}
\end{aligned}
$$

and $P_{\mathrm{XZ}} = P_{\mathrm{YZ}}$ by symmetry. By further marginalisation over X:

$$
\begin{aligned}
P_{\mathrm{Z}}(0) &= P_{\mathrm{XZ}}(0,0) + P_{\mathrm{XZ}}(1,0) = \tfrac{1}{2}\\
P_{\mathrm{Z}}(1) &= P_{\mathrm{XZ}}(0,1) + P_{\mathrm{XZ}}(1,1) = \tfrac{1}{2}
\end{aligned}
$$

Verify that $P_{\mathrm{XZ}} = P_{\mathrm{X}}P_{\mathrm{Z}}$, $P_{\mathrm{YZ}} = P_{\mathrm{Y}}P_{\mathrm{Z}}$ and $P_{\mathrm{XY}} = P_{\mathrm{X}}P_{\mathrm{Y}}$, but $P_{\mathrm{XYZ}} \neq P_{\mathrm{X}}P_{\mathrm{Y}}P_{\mathrm{Z}}$. $\square$

- We define the expectation of a random variable as the "centre of mass" of its distribution:

$$\mathbb{E}[\mathrm{X}] = \sum_{x \in \mathbb{X}} x \, P_{\mathrm{X}}(x) \qquad \text{(OK to write } E[X]\text{)}$$

- More generally, we can define the expectation of *any function* of *any number* of random variables; for example

$$\mathbb{E}[g(\mathrm{X})] = \sum_{x \in \mathbb{X}} g(x) P_{\mathrm{X}}(x) \quad \text{or} \quad \mathbb{E}[f(\mathrm{X}, \mathrm{Y})] = \sum_{x, y \in \mathbb{X} \times \mathbb{Y}} f(x, y) P_{\mathrm{XY}}(x, y)$$

- The expectation is linear:

$$\mathbb{E}[a\mathrm{X} + b\mathrm{Y}] = a\,\mathbb{E}[\mathrm{X}] + b\,\mathbb{E}[\mathrm{Y}] \quad \text{for all } a, b \in \mathbb{R} \times \mathbb{R}$$

proof:
$$\begin{aligned}
\mathbb{E}[a\mathrm{X} + b\mathrm{Y}] &= \sum_{x, y \in \mathbb{X} \times \mathbb{Y}} (ax + by) \, P_{\mathrm{XY}}(x, y) \\
&= a \sum_{x \in \mathbb{X}} x \sum_{y \in \mathbb{Y}} P_{\mathrm{XY}}(x, y) + b \sum_{y \in \mathbb{Y}} y \sum_{x \in \mathbb{X}} P_{\mathrm{XY}}(x, y) \\
&= a \sum_{x \in \mathbb{X}} x \, P_{\mathrm{X}}(x) + b \sum_{y \in \mathbb{Y}} y \, P_{\mathrm{Y}}(y) \quad \text{by marginalisation} \\
&= a\,\mathbb{E}[\mathrm{X}] + b\,\mathbb{E}[\mathrm{Y}] \qquad \square
\end{aligned}$$

UNIVERSITY OF CAMBRIDGE
Department of Engineering

- For two *independent* random variables X and Y:

$$\begin{aligned}\mathbb{E}[\mathrm{X}\,\mathrm{Y}] &= \sum_{x,y\in\mathbb{X}\times\mathbb{Y}} x\,y\,P_{\mathrm{XY}}(x,y) = \sum_{x,y\in\mathbb{X}\times\mathbb{Y}} x\,y\,P_{\mathrm{X}}(x)\,P_{\mathrm{Y}}(y)\\ &= \sum_{x\in\mathbb{X}} x\,P_{\mathrm{X}}(x) \sum_{y\in\mathbb{Y}} y\,P_{\mathrm{Y}}(y)\\ \mathbb{E}[\mathrm{X}\,\mathrm{Y}] &= \mathbb{E}[\mathrm{X}]\,\mathbb{E}[\mathrm{Y}]\end{aligned}$$

- This is not true in general of two random variables that are not independent.
- Two random variables X and Y for which $\mathbb{E}[\mathrm{X}\,\mathrm{Y}] = \mathbb{E}[\mathrm{X}]\,\mathbb{E}[\mathrm{Y}]$ are said to be uncorrelated.
- X and Y independent $\Rightarrow$ X and Y uncorrelated but the inverse is not true.

The expectation can be used to further characterise a probability function by calculating its central moments.

- The central *second* moment, or variance, characterises the "spread" of the distribution as
$$\mathrm{Var}[\mathrm{X}] = \mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])^2]$$
the averaged squared-difference to the mean.
- The central *third* moment $\mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])^3]$, characterises the "asymmetry" of the distribution.
- The central *fourth* moment $\mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])^4]$, characterises the "tailedness" of the distribution.
- The central n-th moment $\mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])^n]$, characterises higher-order features of the shape of the distribution.

We can rewrite the variance as $\mathrm{Var}[\mathrm{X}] = \mathbb{E}[\mathrm{X}^2] - \mathbb{E}[\mathrm{X}]^2$
$$\begin{aligned}\mathrm{Var}[\mathrm{X}] &= \mathbb{E}[\mathrm{X}^2] - 2\mathbb{E}[\mathrm{X}\,\mathbb{E}[\mathrm{X}]] + \mathbb{E}[\mathrm{X}]^2\\ &= \mathbb{E}[\mathrm{X}^2] - 2\mathbb{E}[\mathrm{X}]^2 + \mathbb{E}[\mathrm{X}]^2\\ \mathrm{Var}[\mathrm{X}] &= \mathbb{E}[\mathrm{X}^2] - \mathbb{E}[\mathrm{X}]^2 \qquad \square\end{aligned}$$

- The idea of entropy[6] in information theory was introduced by Claude Shannon (1916-2001) to convey the "surprise" of an event:
  - an event with 100% probability is perfectly unsurprising (and yields no information);
  - the less probable an event is, the more surprising it is (and the more information it yields).
- We define the *information content* of a random variable X as

$$I_{\mathrm{X}}(x) = -\log_2 P_{\mathrm{X}}(x)$$

- The entropy (in *bit*) is the average of the information content:

$$\mathbb{E}[I_{\mathrm{X}}(\mathrm{X})] = \mathbb{H}[\mathrm{X}] = -\sum_{x \in \mathbb{X}} P_{\mathrm{X}}(x) \log_2 P_{\mathrm{X}}(x)$$

(OK to write $\mathrm{H}[\mathrm{X}]$)

- We will discuss this, with examples, in the next part.

[6] we're not talking about thermodynamics here, don't panic

You can attempt Problems 1 to 5 of Examples Paper 5