MAST20004 Probability
Semester 2, 2020

Lecturer: Xi Geng

Why do we learn probability?

- Probability theory is the foundation of statistics.
- It is used to describe mathematical models of real world problems that have a nature of randomness:
  - Economics, finance.
  - Physics, chemistry, biology.
  - Computer science (algorithms, machine learning)
- After Kolmogorov (1930s), probability theory becomes a rigorous branch of mathematics that
  - lies as the foundation of stochastic processes, stochastic calculus etc.
  - provide new ideas and tools to solve problems in many areas of mathematics.

The Monty Hall Problem

- A prize lies behind one of three doors.
- The contestant chooses a door.
- Monty Hall (who knows which door the prize is behind) opens a door not chosen by the contestant that does not have the prize behind. There must be at least one such door.
- Monty Hall then offers the contestant the option of changing his/her original selection to the other unopened door.
- Should the contestant change?

St. Petersburg Paradox

- Toss a fair coin, bet on Tail.
- Bet \$1 in the first toss. End game if I win.
- If I lose, bet \$2 in the second toss. End game if I win.
- If I lose again, bet \$4 in the third toss. End game if I win.
- Keep playing until first Tail appears.
- Should I play this game?

The Bus-Stop Paradox

- Buses on a particular route arrive at randomly-spaced intervals throughout the day.
- On average a bus arrives every hour.
- A passenger comes to the bus-stop at a random instant.
- What is the expected length of time that the passenger will have to wait for a bus?

Monkey Typing Shakespeare

- A monkey types one capital letter randomly at each time.
- Will the monkey eventually produce an exact copy of Shakespeare's "The Tragedy of Hamlet"?
- If yes, how long does it take on average to produce such a copy?

Random Experiments

- *Random experiment:* a process leading to a number (which may be infinite) of possible outcomes and the actual outcome that occurs depends on influences that cannot be predicted beforehand.

- The *sample space* (sometimes also called the *outcome space*), denoted as Ω, is the set of *all* possible outcomes of a random experiment.

Examples

Toss of a coin.

$$\Omega = \{H, T\} \qquad \text{where} \qquad \begin{aligned} H &= \text{“head up”} \\ T &= \text{“tail up”} \end{aligned}$$

Spin of a roulette wheel.

$$\Omega = \{0, 1, 2, \ldots, 36\}$$

(There are 37 numbers on an Australian roulette wheel.)

Remark: Given a random experiment, there may be different ways to define the sample space depending on what we are interested in observing.

For instance, consider a horse race (8 horses with 3 winners).

This is a random experiment because the outcome of the race is not predictable.

If we observe only the winner we might take

$$\Omega = \{\text{all horses in the race}\}$$

since the winner has to be one of the horses. If we observe the placings we could take

$$\Omega = \{\text{all possible ordered sets of } 3 \text{ horses in the race}\}$$

More generally, if we observe the whole race we might take

$$\Omega = \{\text{all possible finishing orders}\}$$

or, even

$$\Omega = \{\text{all possible finishing orders} \\ \text{together with times}\}$$

This example illustrates that a given physical situation can lead to different sample spaces depending on what we choose to observe.

Some Further Examples

- A coin is tossed until a head occurs and the number of tosses required is observed $\Omega = \{1, 2, 3, \ldots\}$.
- A machine automatically fills a one litre bottle with fluid, and the actual quantity of fluid in the bottle is measured in litres $\Omega = \{q : 0 \leq q \leq 1\}$.
- A car is filled up with petrol and then driven until it runs out, the distance it travels is measured in kilometres $\Omega = \{d : 0 \leq d < \infty\}$.

Simulation

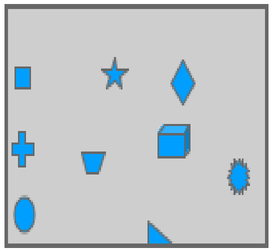
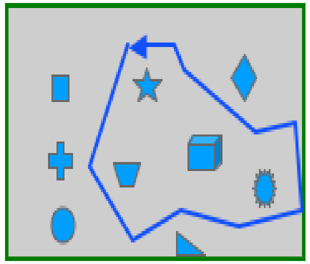
Simulation of random experiments is a tool which probabilists often use. It consists of performing the experiment on a computer, instead of in real life. This has many advantages. For instance:

- It enables us to try out multiple possibilities before building a physical system.
- It is possible to perform multiple repetitions of an experiment in a short time, so that precise estimates of the behaviour can be derived.

In our computer lab classes, we shall be using simulation.

Events

- We are often interested in a group of outcomes.
- An *event* is a set of possible outcomes, that is a subset of Ω.

- We say that the event A occurs if the observed outcome ω of the random experiment is one of the outcomes in the set A (symbolically, $\omega \in A$).

Examples

Toss of a die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$.

The event that "the number on the die is even" is

$$A = \{2, 4, 6\}.$$

Spin of a roulette wheel. The sample space is $\Omega = \{0, 1, 2, \cdots, 36\}$.

The event that "one of the first three numbers occurs" is

$$B = \{1, 2, 3\},$$

and the event that "the number 0 comes up" is

$$D = \{0\}.$$

Since Ω is a set of outcomes, Ω itself is an event. This is known as the *certain event*. One of the outcomes in Ω must occur.

The empty set $\emptyset$ is also an event, known as the *impossible event*.

Remark: In general, when Ω is an infinite set, it is usually *not* feasible to view *every* subset of Ω as a legal event.

- Will cause trouble when assigning probabilities on events (this is a deep theoretical point).

We often need to specify a class of events along with the sample space.

Event Relations and Operations

Events are sets and so they are subject to the normal set operations. For instance:

- The event $A \cup B$ is the event that A *or* B *or* both occur.
- The event $A \cap B$ is the event that A *and* B both occur.
- The event A^c is the event that A does not occur.
- We write $\omega \in A$ to say that the outcome ω is in the event A.
- We write $A \subseteq B$ to say that A is a subset of B. This includes the possibility that $A = B$.
- If A is finite (which will often not be the case), we write $\#A$ for the number of elements of A.

For illustrative purposes, and to gain intuition, the relationship between events is often depicted using a *Venn diagram*.

Two events A_1, A_2 which have no outcomes in common ($A_1 \cap A_2 = \emptyset$) are called *disjoint* (or *mutually exclusive*) events.

Similarly, events $A_1, A_2, \ldots$ are *disjoint* if no two have outcomes in common, that is

$$A_i \cap A_j = \emptyset \quad \forall\, i \neq j.$$

Two events A_1, A_2 are *exhaustive* if they contain all possible outcomes between them,

$$A_1 \cup A_2 = \Omega.$$

Similarly, events $A_1, A_2, \ldots, A_n$ (where n may take ∞) are exhaustive if their union is the whole sample space,

$$\bigcup_{i=1}^{n} A_i = \Omega.$$

Examples

1. Since $A \cap A^c = \emptyset$, A and A^c are disjoint.
2. Since $A \cup A^c = \Omega$, A and A^c are exhaustive.
3. Throw of a die. Let

   $$A = \{1, 3, 5\}, \quad B = \{2, 4, 6\}, \quad C = \{1, 2, 4, 6\}, \quad D = \{2, 4\}.$$

   Then A and B are disjoint and exhaustive, A and C are exhaustive but not disjoint and A and D are disjoint but not exhaustive.

Set operations satisfy the distributive laws

$$
\begin{aligned}
A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\
A \cup (B \cap C) &= (A \cup B) \cap (A \cup C)
\end{aligned}
$$

and De Morgan's laws

$$
(A \cup B)^c = A^c \cap B^c
$$

$$
(A \cap B)^c = A^c \cup B^c
$$

Defining Probability (Ghahramani 1.1)

Up to now we have talked about ways of describing results of random experiments – an event A happens if the outcome of the experiment is in the set A. We haven't yet talked about ways of assigning a measure to the "likelihood" of an event happening.

That is, we are yet to define what we mean by a probability.

First let us think about some intuitive notions.

What do we mean when we say "The probability that a toss of a coin will result in 'heads' is 1/2"?

An interpretation that is accepted by most people for practical purposes, that such statements are made based upon some information about *relative frequencies*.

People	#trials	#heads	frequency of heads
Buffon	4040	2048	0.5069
De Morgan	4092	2048	0.5005
Feller	10000	4979	0.4979
Pearson	12000	6019	0.5016
Pearson	24000	12012	0.5005

Similar statements can be made about tossing dice, spinning roulette wheels, arrivals of phone calls in a given time period, etc.

Hence it seems that we can think of a probability as a long term relative frequency. However there are problems with this interpretation. Consider the statement

"The probability that horse X will win the Melbourne Cup this year is $1/21$".

A similar statement is

"The probability that *macrotis lagotis* will be extinct in 100 years is $1/100$".

Both of the above-mentioned experiments will be performed only once under unique conditions, so a repetitive relative frequency definition makes no sense.

Another way to think of probability in these experiments is that it reflects the odds at which a person is willing to bet on an event.

Thus probability takes on a "personal" definition: my evaluation of a probability may not be the same as yours. This interpretation of probability is known as the *Bayesian interpretation*.

How do mathematicians define probabilities?

Through a set of *axioms* under which probabilities behave "naturally".

What we mean by "naturally" is quite simple:

- We assign the value 1 to be the probability of the certain event and require that the probability of any event be nonnegative.

- If A and B are disjoint events, then the occurrence of A implies B can't happen, and vice versa. Thus we would expect that

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B),$$

where $\mathbb{P}(A)$ denotes the probability of event A.

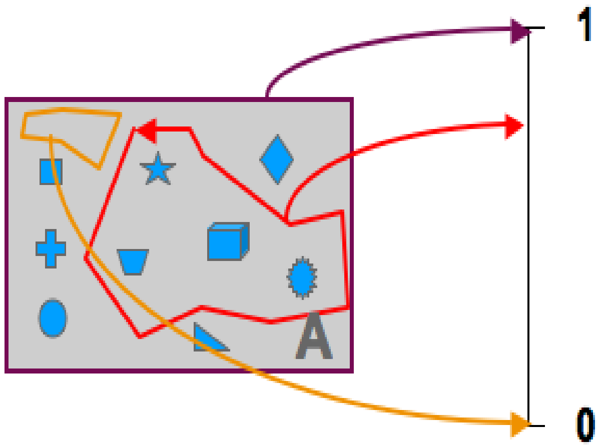
Probability axioms (Ghahramani 1.3, 1.4)

These considerations lead to the following *axioms*:

1. $\mathbb{P}(A) \geq 0$, for all events A.
2. $\mathbb{P}(\Omega) = 1$.

3*. (Finite additivity) For a set of mutually disjoint events $\{A_1, A_2, A_3, \ldots, A_n\}$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$



For both theoretical and practical reasons, we need a slightly stronger version of Axiom 3*. More precisely, we need to require it to hold for *infinite sequences of mutually disjoint events*. Thus, we replace it by the following axiom:

3. (Countable additivity)

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

where $\{A_1, A_2, A_3, \ldots\}$ is any sequence of mutually disjoint events.

Andrey Kolmogorov [25/04/1903 - 20/10/1987]

We use countable, rather than finite, additivity because we sometimes need to calculate probabilities for countable unions.

For example, the event that a 6 eventually occurs when tossing a die can be expressed as $\bigcup_{i=1}^{\infty} A_i$, where A_i is the event that the 6 occurs for the first time on the ith toss.

The gap between Axiom 3* and Axiom 3 is beyond the scope of this subject and it will be discussed in MAST30020 Probability for Inference.

From the axioms, we can deduce the following properties of the probability function:

(4) $\mathbb{P}(\emptyset) = 0$, since $\emptyset \cup \emptyset \cup \cdots = \emptyset$

(5) Finite additivity

(6) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, since $A \cup A^c = \Omega$

(7) $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$, since $A \cup (A^c \cap B) = B$

(8) $\mathbb{P}(A) \leq 1$, since $A \subseteq \Omega$

(9) Addition theorem:

$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

(10) Continuity: either (a) $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ and $B = \cup_{i=1}^{\infty} A_i$, or (b) $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$ and $B = \cap_{i=1}^{\infty} A_i$, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(B)$

Remarks

- $\mathbb{P}(\cdot)$ is a set function. It maps $\mathcal{A} \to [0, 1]$, where $\mathcal{A}$ denotes the class of events, that is the set of subsets of the outcome space.
- For a discrete sample space, we can write

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}).$$

- In general, "possible" outcomes are allowed to have zero probability, thus $\mathbb{P}(E) = 0 \quad \not\Rightarrow E = \emptyset$. Similarly, there can be sets other than Ω that have probability 1.

Evaluating Probabilities

So far we have said nothing about how numerical values are assigned to the probability function, just that if we assign values in such a way that the Axioms (1) – (3) hold, then the properties (4) – (10) will also hold.

Assigning probabilities to events is a large part of what the subject is about.

- There may be no 1 "right" answer!
  - Simple problems may have a single reasonable solution
  - Real life problems often have many possible solutions
    - each OK, if they obey the three axioms
    - selection uses art and science

The Simplest Case
(Classical Probability Model)

The outcome space is finte: $\#(\Omega) = N$.

The class of events $\mathcal{A}$ is the collection of all subsets of Ω.

Assign probabilities to events in the way that all outcomes occur equally likely, that is,

$$\mathbb{P}(\{\omega\}) = 1/N$$

for all $\omega \in \Omega$, and further,

$$\mathbb{P}(A) = \frac{\#(A)}{N}.$$

Example: Coin Tossing

What is the probability of having a "tail" when tossing a fair coin? Sample space:

$$\Omega = \{H, T\}.$$

Define the probability function induced by:

$$\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}.$$

Therefore,

$$\mathbb{P}(\text{having a tail}) = \mathbb{P}(\{H\}) = \frac{1}{2}.$$

Suppose that we toss a pair of fair coins. What is the probability of having a "head" and a "tail"?

This is equivalent to tossing two coins (called A and B) independently. The sample space consists of *ordered* pairs of outcomes (corresponding to Coins A and B):

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}.$$

Define the probability function induced by:

$$\mathbb{P}(\{(H, H)\}) = \mathbb{P}(\{(H, T)\}) = \mathbb{P}(\{(T, H)\}) = \mathbb{P}(\{(T, T)\}) = \frac{1}{4}.$$

Therefore,

$$\mathbb{P}(\text{having a head and a tail}) = \mathbb{P}(\{(H, T), (T, H)\}) = \frac{1}{2}.$$

D'Alembert's Solution (1717-1783)

There are three possibilities when tossing a pair of coins: H&H, H&T, T&T. Therefore, the probability of having a "head" and a "tail" is $1/3$.

What goes wrong with this solution?

Implicitly, the sample space is taken to be the set of *unordered* pairs of outcomes:

$$\Omega = \{\{H, H\}, \{H, T\}, \{T, T\}\}.$$

D'Alembert worked with the classical probability model over Ω:

$$\mathbb{P}(\{\{H, H\}\}) = \mathbb{P}(\{\{H, T\}\}) = \mathbb{P}(\{\{T, T\}\}) = \frac{1}{3}.$$

Theoretically, this is a well-defined probability space. But this model does not practically correspond to the problem! (Try to toss a pair of coins for 30 times by yourself, and observe the number of times having "head + tail").

NB. We often use $(\cdots)$ to denote ordered tuples, and use $\{\cdots\}$ to denote unordered tuples.

The Birthday Problem

In a group with n people, what is the probability that at least two of them have the same birthday?

Suppose that the n people are independent, and any day of the year is equally likely to be the birthday of a person.

The sample space consists of ordered n tuples in which the i-th component records the possible birthday of the i-th person:

$$\Omega = \{(d_1, \cdots, d_n) : 1 \leqslant d_i \leqslant 365, 1 \leqslant i \leqslant n\}.$$

The problem is described by the classical probability model over Ω: for each $(d_1, \cdots, d_n) \in \Omega$,

$$\mathbb{P}(\{(d_1, \cdots, d_n)\}) = \frac{1}{365^n}.$$

Define

$$A: \text{ at least two people have the same birthday.}$$

To compute $\mathbb{P}(A)$, it is easier to compute the probability of its complement:

$$A^c: \text{ all } n \text{ people have different birthdays}$$

$$\#(A^c) = 365 \times 364 \times 363 \cdots \times (365 - n + 1) \qquad \text{(why?)}$$

$$\begin{aligned}\mathbb{P}(A^c) = \frac{\#(A^c)}{365^n} &= \frac{365 \times 364 \times 363 \cdots \times (365 - n + 1)}{365^n} \\ &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \cdots \times \left(1 - \frac{n-1}{365}\right).\end{aligned}$$

A trick to estimate the product numerically:

$$
\begin{aligned}
&\left(1-\frac{1}{365}\right)\times\left(1-\frac{2}{365}\right)\cdots\times\left(1-\frac{n-1}{365}\right)\\
&=\exp\left(\sum_{k=1}^{n-1}\log\left(1-\frac{k}{365}\right)\right)\\
&\approx\exp\left(-\sum_{k=1}^{n-1}\frac{k}{365}\right) \qquad (\log(1+x)\approx x \text{ when } x \text{ small})\\
&=\exp\left(-\frac{1}{365}\times\frac{n(n-1)}{2}\right).
\end{aligned}
$$

Therefore,

$$\mathbb{P}(A)=1-\mathbb{P}(A^c)\approx 1-\exp\left(-\frac{1}{365}\times\frac{n(n-1)}{2}\right).$$

In a group with $n = 23$ people, the probability that at least two people have the same birthday is approximately 0.5!