Probability Theory

Aiphabet - aiphabet.org

Uncertainty is all around us! Most events in life involve some uncertainty, from our chances to succeed in a career to our luck finding a parking spot. So, how can we express events involving uncertainty, likelihood, risk, or chance? The theory of probability provides a rigorous framework to reason about uncertainty, quantify it, and study the laws that govern chance. It is thanks to probability that much of science, engineering, and other areas have made significant progress today. An example is artificial intelligence, where probability allowed the field to make a huge leap by modeling uncertainty in applications as diverse as machine translation and robotics.

Prerequisites: Set theory, functions, and counting.

1 Basic Definitions

Probability starts with the central idea of random experiment and sample space.

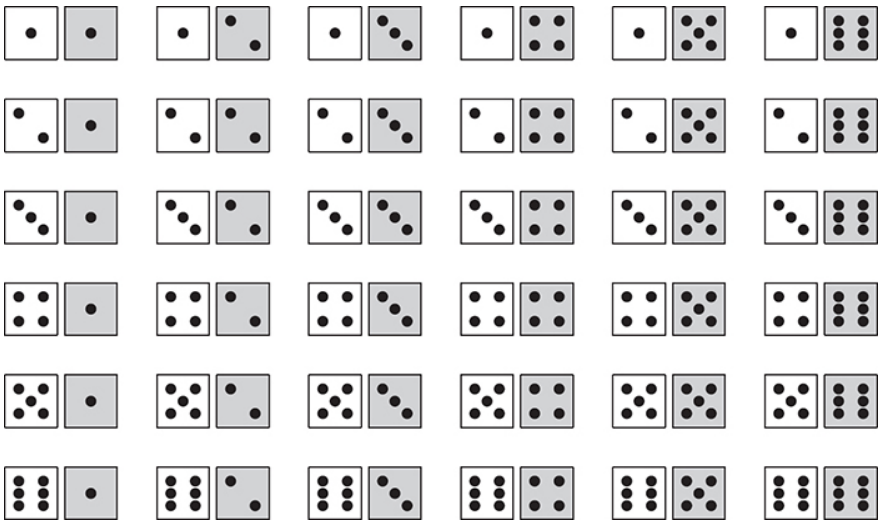
Definition 1 ***Random Experiment:*** *A random experiment consists of a process whose outcome cannot be predicted with certainty, but the set of possible outcomes is known.*

Definition 2 ***Sample space:*** *The sample space S of an experiment is the set of all possible outcomes of the experiment.*

Example 1

Here are the possible outcomes of random experiments:

1. The flip of a fair coin: $\{H, T\}$
2. The flip of two fair coins: $\{(H, H), (H, T), (T, H), (T, T)\}$; for simplicity, we also write: $\{HH, HT, TH, TT\}$.
3. The roll of a die: $\{1, 2, 3, 4, 5, 6\}$
4. The roll of two distinguishable dice: $|S| = 6 \times 6 = 36$. The sample space is:



Definition 3 ***Probability function:*** *A probability function P takes an outcome $s \in S$ and return the probability of s, denoted*

$$P : S \longrightarrow [0, 1]$$

such that:

$$0 \leq P(s) \leq 1 \ \ \forall s \in S$$

$$\sum_{s \in S} P(s) = 1$$

Assigning probabilities

How are probabilities assigned? We will assume S is finite and that all outcomes are equally likely.

The probability of an outcome $s \in S$ is given by:

$$\forall s \in S \quad P(s) = \frac{1}{|S|}$$

Example:

1. We toss a fair coin one time. The sample space is:

   $$S = \{H, T\}$$

   Because it is a fair coin (the weight of the coin is uniformly distributed), it has 50% chance of landing on Heads and 50% chance of landing on Tails.

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

2. We toss a fair coin 3 times in a row. The sequence of Heads and Tails is recorded. The sample space is:

   $$S = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$$

   There are $2^3 = 8$ such lists. In other words $|S| = 8$ $P(s) = \frac{1}{8}$ $\forall s \in S$

2 Events

Instead of individual outcomes, we might be interested in a subset of the sample space with some characteristics. In the example above with

$$S = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\},$$

an event could be "Getting exactly two Tails in flipping a fair coin three time in a row." The event is then the subset of all outcomes fulfilling the event description. In this case, this subset is $\{TTH, THT, HTT\} \subseteq S$

Definition 4 ***Event:*** *An event is any subset A of the sample space S.*

$$A \subseteq S$$

Definition 5 ***Probability of an event:*** *The probability of A, denoted $P(A)$ is the sum of the probabilities of all outcomes that belong to A.*

$$P(A) = \sum_{a \in A} P(a)$$

Note:
$0 \leq P(A) \leq 1 \quad \forall A \subseteq S$

Example: We toss a coin five times in a row.

1. Let A denotes the event that exactly one Heads emerges.

$$A = \{HTTTT, THTTT, TTHTT, TTTHT, TTTTH\}$$

A contains 5 outcomes, each of which has a probability of $\frac{1}{32}$. $P(A) = \frac{5}{32}$

Axioms of probability

Let S be a sample space space and let A and B be two events. The following properties hold:

1. $P(\emptyset) = 0$
2. $P(S) = 1$
3. $P(\bar{A}) = 1 - P(A)$ where $\bar{A}$ denotes the complement of A.
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. $P(A \cup B) \leq P(A) + P(B)$
6. If $A \cap B = \emptyset$ (A and B are disjoint) then $P(A \cup B) = P(A) + P(B)$

7. If $A \subseteq B$ then $P(A) \leq P(B)$

3 Conditional probability

Definition 6 Conditional Probability: *Let S be a sample space and let P be a probability over S. Let A and B be events such that $P(B) > 0$. The conditional probability of A given B is given by:*

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

We can think of conditional probability as a restriction of the set of possible outcomes from the sample space S to B.

Example 2

If we roll two fair dice and observe that the first die is a four, what is the probability that the sum of the two dice equals a six **given that the first die is a four**?

Given that the first roll is a 4, the restricted sample space is $\{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}$. The conditional probability of this event is 1/6. Another way to calculate it is to use the formula $\frac{1/36}{6/36} = 1/6$.

4 Marginals and Total Probability Rule

A useful way to organize the probability space of two events A and B is to divide the sample space into four mutually exclusive events (note: the logical motivation here is similar to the concept of a proof by cases: we attack each independent case separately, and put them all together at the end of present a single solution).

	A	$\bar{A}$	
B	P($A \cap B$)	P($\bar{A} \cap B$)	P(B)
$\bar{B}$	P($A \cap \bar{B}$)	P($\bar{A} \cap \bar{B}$)	P($\bar{B}$)
	P(A)	P($\bar{A}$)	1

Marginals

The probability in the margins are called marginals and are calculated by summing across the rows and the columns. The probability of two events $A \cap B$ is called joint distribution. Sometimes we denote $A \cap B$ as A, B or AB.

From the marginals and the definition of conditional probability, we have:

$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

Definition 7 Total Probability Rule*:*

$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

It is called **Total** *because* A *and* $\overline{A}$ *form the totality of the sample space.*

Using conditional probability we can write:

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

5 Bayes rule

We can derive a useful rule, called **Bayes Rule** (after the English philosopher Thomas Bayes by using the definition of conditional probabilities:

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

$$P(A|B)P(B) = P(B|A)P(A)$$

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)}$$

$P(A|B)$ is called posterior (posterior distribution on A given B.)
$P(A)$ is called prior.
$P(B)$ is called evidence.
$P(B|A)$ is called likelihood.
This formula known as *Bayes rule*.
Using the table above, we can write $P(A|B)$ as follows:

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

$$P(A \cap B) = P(B \cap A) = P(B|A)P(A)$$

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

$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

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

A common use of Bayes rule is when we want to know the probability of an unobserved event given an observed event.

6 Intersection of events

We often like to calculate the joint probability of events; this will depend on whether they are independent or not:

Definition 8 *Two events A and B are independent provided that their joint distribution is the product of their marginal distributions:*

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

We denote A and B independent as follows: $A \perp\!\!\!\perp B$

Another way to see this is to reorder the terms in the formula for conditional probability (as probability of event A is not affected by event B, i.o.w. $P(A \mid B) = P(A)$):

$$P(A|B)P(B) = P(A \cap B)$$
$$P(A|B)P(B) = P(A)P(B) = P(A \cap B)$$

Note:
Probability of intersection of mutually independent events is the product of their probabilities.

Example 1:
We toss a fair coin three times. Consider the following events:

- A_1: Event of obtaining Tails in the first toss.
- A_2: Event of obtaining Tails in the second toss.
- A_3: Event of obtaining Heads in the third toss.

1. What is the probability of A1, A2, and A3 happening?

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3)$$

Since it is a fair coin:

$$P(A_1 \cap A_2 \cap A_3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$$

Checking independence
To check two events are independent, just check that the joint probability is equal to the product of the marginal probabilities.
For instance: $P(hot, sun) = 0.4$ $P(hot) = 0.5$ $P(sun) = 0.5$ $P(hot, sun) \neq P(hot) \times P(sun)$.

If the events are not mutually independent, how do we calculate their joint probability? From the conditional probability formula, we have:

$$P(A \cap B) = P(A)P(B|A)$$

We can extend it to three events, as follows:

$$P(A \cap B \cap C) = P(A \cap B)P(C|A \cap B) = P(A)P(B|A)P(C|A \cap B)$$

We call this formula the chain rule that is very useful to estimate the joint probability of non independent events in experiments involving a sequence of choices. We generalize the rule as follows:

Definition 9 (Chain rule): *For any events* A_1, A_2, $\cdots$, A_n*:*

$$P(A_1 \cap A_2 \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|\bigcap_{i=1}^{n-1} A_i)$$

An alternative way to write it is:

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

Conjunction/Intersection of events

The intersection of events is equivalent to the notion of the conjunction. In other words, $A_1 \cap A_2$ means event A_1 happened and event A_2 happened. To express the disjunction between events, we will use the union, as defined in the next section.

7 Union of events

Events are sets. So, it follows that theorems or principles that apply to sets, such as the principle of inclusion-exclusion (PIE), which we used to calculate the size of unions of sets, also apply to events.

Here we are counting the size of the union proportionally to the size of the sample space.

$$P(A_1 \cup A_2 \cup A_3) = P(A_1)+P(A_2)+P(A_3)-P(A_1 \cap A_2)-P(A_1 \cap A_3)-P(A_2 \cap A_3)+P(A_1 \cap A_2 \cap A_3)$$

More generally:

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) + (-1) \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} P(\bigcap_{i=1}^{n} A_i)$$

8 Conditional independence

We have seen that if all variables are mutually independent, it is easy to calculate the joint distribution (e.g. n coin flips) by multiplying the individual probabilities. However, unconditional independence is rare, and unrealistic. It is also useless. some variables are easy to measures and others are hidden (latent) and we hope we can measure them. If everything is independent of everything else than we can't measure the latent variables.

We also saw that if variables are not mutually independent, we can use the chain rule to calculate the joint distribution.

What if don't have full Independence but a set of variables is independent of another sets of variables? We can exploit this partial independence that we call conditional independence to simplify the calculations.

Definition 10 (Conditional independence) *Let X, Y, and Z be random variables. We say that X is conditionally independent of Y given Z provided the probability distribution governing X is independent of the value of Y given the value of Z; that is:*

$$(\forall x, y, z)\ P(X = x|Y = y, Z = z) = P(X = x|Z = z)$$

more compactly, we write

$$P(X|Y, Z) = P(X|Z)$$
$$P(Y|X, Z) = P(Y|Z)$$

or equivalently:

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

We write:

$$X \perp\!\!\!\perp Y|Z$$

What happens to the chain rule in the case of conditional independence?

$$P(Z, X, Y) = P(Z)P(X|Z)P(Y|Z, X)$$

Given that:

$$X \perp\!\!\!\perp Y|Z$$

Then we simplify the chain rule as follows:

$$P(Z, X, Y) = P(Z)P(X|Z)P(Y|Z)$$

Conditional probability is a key concept in graphical models such as Bayes Nets, that encodes probabilistic relationships among variables.

9 Practice Problems

1. In a reliability test, a light switch is turned on and off until it fails. If the probability that the switch will fail every flip is 0.001, what is the probability that the switch will fail after exactly 1200 flips?

2. Three people are selected at random. Assume there are 365 days in a year.

   (a) What is the probability that all three people have the same birthday?

   (b) What is the probability that none of the three have the same birthday?

3. Freddie picks a number in the from 1 to 2023. At the same time, Toby picks a number from 1 to 4046. Assume that they have picked different numbers. What is the probability that Toby has picked a number higher than Freddie's?

9.1 Solutions

1. In order for the light switch to fail after exactly 1200 flips, it must have functioned successfully 1200 times first, and then fail right afterwards. Thus, our answer is $(1 - 0.001)^1200 * 0.001$.

2. (a) Let's select any day X for the first person in our group's birthday. The probability that the second person's birthday is X is $\frac{1}{365}$, and the probability that the third person's birthday is X is also $\frac{1}{365}$. Thus, the total probability is $1 * \frac{1}{365} * \frac{1}{365} = \boxed{\frac{1}{133225}}$.

   (b) Once again, Let's select any day X for the first person in our group's birthday. The probability that the second person's birthday is not X (call it Y) is $\frac{364}{365}$, as there are 364 days left to choose from. Using similar logic the probability that the third person's birthday is unique is $\frac{363}{365}$, as there are 363 days to choose from that are not X or Y. This leaves our final answer as $1 * \frac{364}{365} * \frac{363}{365} = \frac{363*364}{365^3} = \boxed{\frac{132132}{133225}}$.

3. There is a $\frac{1}{2}$ chance that Toby picks a number from 2024 to 4046. In this case, Toby will always have picked a number higher than Freddie, because the largest number Freddie can pick is 2023. The other half of the time, both Toby and Freddie have picked a number from 1 to 2023. In this case, each has a $\frac{1}{2}$ chance of picking a number higher than the other. Thus, the total probability that Toby has picked a higher number than Freddie is $\frac{1}{2} * 1 + \frac{1}{2} * \frac{1}{2} = \boxed{\frac{3}{4}}$