Probability

An experiment that has many possible outcomes. We have tossed a coin many times, and this is what we have found. Each toss produces either a head or a tail, but we cannot predict which one will be produced before the toss. This experiment shares with many other experiments several salient features, which we next paraphrase in the language of sets.

A possible outcome of an experiment is called a *sample point*. The set of all possible outcomes of an experiment is called the *sample space*. A subset of possible outcomes is called an *event*.

We label the outcomes of an experiment by $\gamma_1, \gamma_2, ...\gamma_i, ...$ The subscripts differentiate the outcomes of the experiment, but do not imply any order among them. The set of all possible outcomes of the experiment is the sample space Γ:

$$\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_i, ...\}.$$

Here are some subsets of outcomes:

$$A = \{\gamma_1, \gamma_2, \gamma_3\},\ B = \{\gamma_2, \gamma_3, \gamma_4\},\ C = \{\gamma_2, \gamma_3\},\ D = \{\gamma_7, \gamma_8, \gamma_9\}.$$

The set *A* is the event that either γ_1 or γ_2 or γ_3 occurs. The point γ_1 belongs to the event *A*, $\gamma_1 \in A$, but does not belong to the event *B*, $\gamma_1 \notin B$. The *union* of two sets, $A \cup B = \{\gamma_1, \gamma_2, \gamma_3, \gamma_4\}$, is the event that either γ_1 or γ_2 or γ_3 or γ_4 occurs. The *intersection* of two sets, e.g., $AB = \{\gamma_2, \gamma_3\}$, is the event that either γ_2 or γ_3 occurs. The event *C* is a subset of event *A*, namely, $C \subset A$; consequently, if event *C* occurs, event *A* must also occur. The event *D* shares no sample points with event *A*, namely, $A \cap D$ = empty set; the events *A* and *D* are said to be *mutually exclusive* or *disjoint*.

Examples. *(a) Tossing a coin once.* The sample space of tossing a coin once is the set $\{H, T\}$. Here are some events and their corresponding sets:

- the experiment produces a head, $\{H\}$;
- the experiment produces either a head *or* a tail, $\{H\}\cup\{T\}$, which is the sample space;
- the experiment produces both a head *and* a tail, $\{H\}\{T\}$, which is an empty set;
- the experiment produces neither a head nor a tail, which is also an empty set.

(b) Tossing a coin twice. The sample space for tossing the coin twice is the set $\{HH, HT, TH, TT\}$. Some possible events are:

- the experiment produces two heads, $\{HH\}$;
- the experiment produces exactly one head, $\{HT, TH\}$;
- the experiment produces at least one head, $\{HH, HT, TH\}$;
- the experiment produces no head, $\{TT\}$.

(c) Throwing a die. A die thrown once produces one of six possible outcomes, depending on which face is on the top. The sample space of throwing a die once is the set $\{1,2,3,4,5,6\}$. The sample space of throwing a die twice is the set

$$\begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}$$

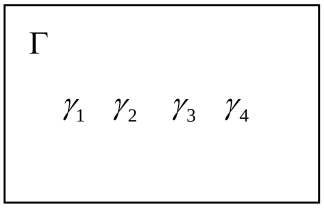
(d) Throwing both a coin and a die once. The sample space of throwing both a coin and a die once is the set

$$\{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}.$$

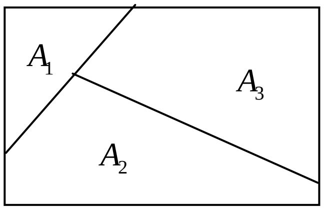
(e) A physical system. Consider a physical system, such as a pile of molecules or a cavity of vacuum. The system has many quantum states, and does an “experiment” all by itself: the system rapidly switches from one quantum state to another. Each quantum state of the system is a sample point of the experiment.

Construct a sample space at a suitable level of detail. The sample space is a special event, so is every sample point. Also, we can regard an event as a sample space: a subset may be considered as a set in its own right; we speak of a subset merely to indicate that we have a larger set in the back of our minds. Furthermore, we can divide a sample space into a class of disjoint events, and then regard each event as a sample point.

This last change of view is particularly important because individual outcomes of an experiment are often too numerous to interest us; we would rather focus on some aggregates of outcomes. When tossing a coin, a gambler is interested in whether the outcome is a head or a tail. Of course, the coin is made of atoms, each atom is made of electrons, protons and neutrons, and each neutron is made of… Associated with a head, the coin has numerous quantum states. But this mode of description is too detailed to interest the gambler, unless she happens to be a physicist.



A sample space



A dissection

Use a class of disjoint events to dissect a sample space. Denote the sample space of an experiment by $\Gamma = \{\gamma_1, \gamma_2, ...\}$. By a *dissection* of the sample space is meant a class $\mathcal{A} = \{A_1, A_2, ..., A_i, ...\}$ of disjoint events whose union is the sample space.

Example. *The conformations of an RNA molecule.* A system consists of a short RNA molecule in a liquid. The RNA molecule can be in two *conformations*: chains or loops. Each conformation is a gross description consisting of many quantum states. For example, both a straight chain and a wiggled chain belong to the conformation of chains. Even when the shape of the RNA molecule is fixed, the molecules in the surrounding liquid can take many configurations. The two conformations can be readily differentiated by biophysical methods. By contrast, the quantum states may be too numerous to interest us.

In this example, all the quantum states constitute a sample space. The two conformations are events. The two events are disjoint and their union is the sample space, so that the two conformations form a dissection of the sample space of all the quantum states of the system.

How to assign a probability to an event? When a fair coin is tossed once, the two possible outcomes are equally probable, so that the probability of the head and that of the tail are both ½.

But how do we know that a coin is fair? Or, if the coin is biased, how do we assign the probability of each side? To answer these questions, we may toss the coin repeatedly. In general, we repeat an experiment by an ensemble of *n* trials. Say point γ_i occurs n_i times. After a great many trials, if the ratio n_i / n approaches a constant, we call the ratio the probability of sample point γ_i:

$$P(\gamma_i) = n_i / n .$$

A cheater might fabricate a coin with one side lighter than the other. Alternatively, he might be so skilled in tossing to favor head or tail at will. However, for his repeated tosses to constitute a *useful ensemble*, even the cheater must follow a rule: he must toss the coin repeatedly under the same conditions. How do we know that he tosses the coin repeatedly under the same conditions? We proceed as follows. Assuming that he tosses the coin repeatedly under the same conditions, we can make theoretical predictions, and then test them against the record of his tosses. For example, under the assumption, we predict that the ratio $n(H)/n$ approaches a constant after he has tossed the coin many times.

Now suppose that the record shows that $n(H)/n$ indeed approaches a constant. Will we trust him now? Not yet. The cheater may keep most tosses under the same conditions, but does some trick with a few tosses, a cheat that would leave the ratio $n(H)/n$ unchanged when n is large. The theory of probability allows us to make other predictions; for example, given the probability of head, we can predict the probability of an event like "in every 10 tosses, head appears 5 times". We can test such a prediction against the record of his tosses also. The more predictions agree with the record, the more we trust that he tosses the coin under the same conditions. Furthermore, if we use the probability obtained from the record to predict his future performance, we will have to assume that the cheater will throw the die under the same conditions, and make theoretical predictions, and test against his new record, until we find disagreement.

Example. *A physical system kept isolated for a long time.* Even if a system is isolated from the rest of the world, the system is not static: it perpetually switches from one quantum state to another. After the system is isolated for a long time, the probability for the system to be in each state becomes independent of time. All our experience is consistent with the postulate that a system isolated for a long time is equally probable to be in any of its quantum state, a postulate upon which the entire statistical mechanics rests.

Probability of an event. However we assign probabilities to events and then interpret the world, there is a middle step: from known probabilities of some events, we calculate probabilities of some other events. We next prescribe rules for such calculations.

Probability is a *map* from a set of events to a set of nonnegative numbers. Denote the probability of event *A* by $P(A)$. We postulate that the probability follows the following rules:

(a) *P*(sample space) = 1;

(b) $P(A \cup B) = P(A) + P(B)$ for any two disjoint events *A* and *B*.

Let us label all the sample points of an experiment by $\gamma_1, \gamma_2, ...$ A sample point is a special event. Denote the probability of sample point γ_i by $P(\gamma_i)$. The sum of the probabilities of all the sample points is unity:

$$P(\gamma_1) + P(\gamma_2) + ... = 1.$$

Once the probabilities for individual points are known, we can calculate the probability of any event by

$$P(A) = \sum P(\gamma_i).$$

The sum is taken over the sample points that constitute the event *A*.

Conditional probability. Consider a large number of computers shipped from several cities in Asia: Hong Kong, Taiwan, Shanghai, etc. Denote the set of all the computers by Γ, the set of defective computers by *D*, and the set of computers from Hong Kong by *H*. Thus, DH is the set of computers both defective and from Hong. Of the total number of *N* computers, N_H computers come from Hong Kong, and N_{DH} computers from Hong Kong are defective. If we pick one computer randomly from the *N* computers, the probability that the computer is from Hong Kong is $P(H) = N_H / N$, and the probability that the computer is both defective and from Hong Kong is $P(DH) = N_{DH} / N$. The event that a computer is defective given that it is from Hong Kong is denoted by $D|H$. The probability of this event is $P(D|H)$ = N_{DH} / N_H.

In general, let events *H* and *D* be subsets of a sample space Γ, and $P(H) > 0$. Define the probability of event *D* conditional on event *H* as

$$P(D|H) = \frac{P(DH)}{P(H)}.$$

In the above example, by $P(H)$ we mean the probability that a computer is from Hong Kong given that it is from Asia. We ought to write $P(H)$ as $P(H|\Gamma)$. Similarly, we ought to write $P(DH)$ as $P(DH|\Gamma)$. Thus, every probability is conditional on some event.

The above equation may be written as

$$P(DH|\Gamma) = P(D|H)P(H|\Gamma).$$

This equation can be interpreted as picking a computer in two steps: first from the cities in Asia pick a particular city (Hong Kong), and then from the computers from this city pick a particular computer. The probability that a computer is both defective and from Hong Kong given that the computer is from Asia, $P(DH|\Gamma)$, is the probability that a computer is from Hong Kong given that the computer is from Asia, $P(H|\Gamma)$, times the probability that a computer is defective given that the computer is from Hong Kong, $P(D|H)$.

We usually drop the reference to the sample space, and write the above equation as

$$P(DH) = P(D|H)P(H).$$

The two events in $P(DH)$ play symmetric roles, which can be switched. Thus,

$$P(D|H)P(H) = P(H|D)P(D).$$

We have already interpreted the left hand side of this equation. The right hand side means we determine $P(DH)$ by choosing a computer in by two steps via an alternative route. First, we determine the probability that a computer is defective given that the computer is from Asia.

Second we determine the probability that a computer is from Hong Kong given that the computer is defective.

Statistically independent events. When a fair coin is tossed once, the probability to get the head is ½. When a fair die is thrown once, the probability to get face 5 is 1/6. If the two events are independent, the probability to get the head of the coin *and* face 5 of the die is 1/12. We can also view the composite of the two experiments as a single experiment. The composite has 12 sample points, which are equally probable if the coin and the die are independent, so that the probability of each sample point of the composite is 1/12.

In general events *A* and *B* are said to be *statistically independent* if the probability for both events to occur is the product of their individual probabilities, namely,

$$P(AB) = P(A)P(B).$$

Random variable. A random variable is a *map* from a sample space to a set of numbers. Let the sample space of an experiment be $\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_i, ...\}$, and the random variable be *X*. When the experiment produces an outcome γ_i, the random variable takes the value $X(\gamma_i)$. The *domain* of the map is the sample space, which consists of all the outcomes of an experiment, objects that are usually not numbers. The *range* of the map is a set of numbers, which obey usual rules of algebra, such as addition and multiplication. What is random is the outcomes of the experiment; once the outcome is known, the value of the variable is known and not random.

Examples. *(a) Tossing a coin once*. We can assign a random variable *X* for the experiment of tossing a coin. When a toss produces a head, the random variable is one. When a toss produces a tail, the random variable is zero. Thus,

$$X(H) = 1, \qquad X(T) = 0.$$

(b) Tossing a coin seven times. Tossing a coin seven times produces sequences like *HTHTTTH*. A total of $2^7 = 128$ distinct sequences constitute the sample space of this experiment. Let *K* be the number of heads in a sequence; for example, $K(HTHTTTH) = 3$. This experiment is a composite of seven independent tosses. Let X_i be the random variable defined above for the *i*th trial. Thus, for the sequence *HTHTTTH*, these random variables take the following values:

$$X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0, X_5 = 0, X_6 = 0, X_7 = 1 .$$

For any sequence, the random variable *K* is the sum of these random variables:

$$K = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 .$$

(c) Human characteristics. For example, let the sample space be a population of people $\Omega = \{\gamma_1, \gamma_2, ..., \gamma_i, ...\}$, where γ_i is an individual person. Not all human characteristics can be reduced to numbers, but many can. For example, for each individual person γ_i, we can record age $A(\gamma_i)$, weight $W(\gamma_i)$, and height $H(\gamma_i)$. So long as the choice of an individual is random, these characteristics are random variables.

Use a random variable to specify an event. One way to specify an event is to specify sample points by placing a restriction on a random variable. Here are some examples:

- All people of 40 years old, $\{\gamma_i | A(\gamma_i) = 40\}$;
- All people between 100 to 160 pounds, $\{\gamma_i | 100 < W(\gamma_i) < 160\}$;
- All people of 40 years old and between 100 to 160 pounds, $\{\gamma_i | A(\gamma_i) = 40\} \cup \{\gamma_i | 100 < W(\gamma_i) < 160\}$.

Use a random variable to dissect a sample space. A random variable is a map from a sample space to a set of number. This map is often many-to-one, rather than one-to-one. The sample space can be divided into a family of events such that each event consists of all the sample points having the same value of the random variable. These events are mutually exclusive, and their union is the sample space. For example, in a population of people, a subpopulation of people are 40 years old, and another subpopulation of people are 41 years old, etc. Thus, the age of a person can be used to divide the population into a family of subpopulations.

Probability distribution of a random variable. Let X be a random variable defined on a sample space $\Gamma = \{\gamma_1, \gamma_2, ...\}$. Let the range of *X* be the set $\{x_1, x_2, ...\}$. Because a random variable is usually a many-to-one map, the number of distinct values of the random variable is fewer than the number of sample points. For example, if X is a constant *c* for all sample points, the range of X is the set of one number $\{c\}$. Let $P(x_i)$ be the probability of the event that *X* takes the value x_i.

Mean of a random variable. Let $\Gamma = \{\gamma_1, \gamma_2, ...\}$ be the sample space of an experiment, and $P(\gamma_1), P(\gamma_2), ...$ be the corresponding probabilities of the sample points. When the outcome of the experiment is γ_i, a variable *X* takes value $X(\gamma_i)$. Define the *mean* or the *expectation* of the random variable *X* by

$$\langle X \rangle = \sum P(\gamma_i) X(\gamma_i).$$

The sum is taken over all sample points. The mean is a map, whose domain is a set of functions defined on the sample space, and whose range is a set of numbers.

We can divide the sample space into a class of events *A*, *B*,… such that, within each event, every sample point has the same value of *X*. These events are disjoint, and their union is the sample space. Consequently, the class of the events constitutes a dissection of the sample space. Let $X(A)$ be the value of *X* when event *A* occurs, and $P(A)$ be the probability for event *A* to occur. The mean of *X* is

$$\langle X\rangle = P(x_1)x_1 + P(x_2)x_2 + ...$$

This sum is taken over all events in the family.

Variance of a random variable. Define the *variance* of a random variable *X* by

$$Var(X) = \sum P_i(X_i - \langle X\rangle)^2 .$$

The sum is taken over all sample points. The calculation of fluctuation is often aided by the following identity:

$$Var(X) = \langle X^2\rangle - 2\langle X\rangle\langle X\rangle + \langle X\rangle^2 = \langle X^2\rangle - \langle X\rangle^2 .$$

A dimensionless measure of the fluctuation of a random variable. The unit of the mean is the same as that of the random variable. The unit of the variance is the square of that the random variable. The fluctuation of the random variable *X* can be measured by a dimensionless ratio

$$\frac{\sqrt{Var(X)}}{\langle X\rangle} .$$

If we double the random variable, the mean will double and the variance will quadruple, but the fluctuation remains the same.

Example. Let $X_i = i^2$, where *i* is the number on a face of a fair die. Thus,

$$\langle X\rangle = \frac{1^2}{6} + \frac{2^2}{6} + \frac{3^2}{6} + \frac{4^2}{6} + \frac{5^2}{6} + \frac{6^2}{6} = \frac{91}{6}$$

$$\langle X^2\rangle = \frac{1^4}{6} + \frac{2^4}{6} + \frac{3^4}{6} + \frac{4^4}{6} + \frac{5^4}{6} + \frac{6^4}{6} = \frac{2275}{6} ,$$

$$Var(X) = \langle X^2\rangle - \langle X\rangle^2 = \frac{2275}{6} - \left(\frac{91}{6}\right)^2 = 149.08 .$$

$$\frac{\sqrt{Var(X)}}{\langle X \rangle} = 0.81 .$$