A Note on Probability and Typicality

version 0.9

Jan. 20, 2020

Contents

1 What is probability?

Consider two possible events ω_1 and ω_2, one of which is about to happen with probabilities p_1 and the other with p_2. Flipping a coin is an example with ω_1=*head* and ω_2=*tail*. Usually we assume that $p_1 = p_2 = 0.5$ (equal *a priori* probability) for unbiased coin. However, the coin could be biased. So, we dot not assume the *a priori* probability. What does the probability tell us about the event that we will happen in a moment?

Case 1: If $p_1 = 1$ (implies $p_2 = 0$), then we are sure that ω_1 happens and that ω_2 will never happens. There is no uncertainty in this case. Even when there are more than two events, this interpretation is always valid since all other probabilities are zero.

Case 2: If $p_1 = 0$ (implies $p_2 = 1$), then we are sure that ω_1 will not happen. Since there is only two choices, ω_2 must happen for sure in the present situation. However, when there are more than two events, $p_1 = 0$ only means that the corresponding event ω_1 won't happen. It does not tell you what will happen with certainty since we don't know the value of other probabilities, hence uncertainty remains.

Case 3: If $p_1 = 0.5$ (implies $p_2 = 0.5$), we have absolutely no idea which one happens.

Case 4: If $p_1 = 0.8$ (implies $p_2 = 0.2$), we again do not know which happens with any certainty. We think that ω_1 is more likely but what "likely" means?. Since we cannot tell which one happens for sure, it seems that the situation is similar to the case 3. So, what do these two numbers mean to us? That has been a big question and still is. (Check Bayesian inference.)

1.1 Ensemble interpretation

Suppose that there are N identical systems, each of which generates event ω_1 or ω_2 with corresponding probability p_1 and p_2, respectively. They are all independent to each other, meaning that the outcome of each system is not influenced by other systems. In N_1 systems, ω_1 happened and in the remaining N_2 systems ω_2 occurred. Assuming that $N \gg 1$, we can say that $N_1 \approx p_1 N$ and $N_2 \approx p_2 N$. Note that these are not exact equalities. A possible exact statement is

$$p_i = \lim_{N\to\infty} \frac{N_i}{N} \tag{1}$$

which is known as Borel's law of large numbers. This interpretation is not useful when we have only one or few systems. The Borel's law of large numbers is mathematically exact but in practice infinitely large number of systems are not available. In many cases, we have only one system! When weather forecast says the chance of rain is 50%, what does the percentage mean? Will a half of us have rain and the other half have sunshine?

1.2 Information interpretation

We consider again a single system consisting of two events ω_1 and ω_2. Before an event happens, the state of the system is specified by $\{p_1, p_2\}$. Now suppose that ω_1 happened, then the state of the system is changed to $\{1, 0\}$. After the event, uncertainty disappeared. For example, before flipping a coin, the outcome is uncertain. However, once we know it is head. Then, the uncertainty vanishes. The loss of uncertainty corresponds to the gain of information. The amount of information obtained is given by $I_1 = -\log_2 p_1$ measured in *bits*. If the coin is not biased, $p_1 = \frac{1}{2}$ and thus we gain 1 bit of information. This interpretation of probability seems very useful in understanding statistical mechanics. In fact, $I_i = -\log_2 p_i$ is directly related to the concept of entropy which we discuss later. In this interpretation, probability is replaced with equivalent concept *information*. We haven't solved a real problem yet. If you ask ask what is information, the answer is probability! This is tautology.

2 Mathematical definition

Although we don't have a concrete idea of probability, we have an axiomatic foundation of probability theory developed by Andrey Kolmogorov. Mathematical definition of "probability" involves three elements, *sample space*: Ω, *event space*: $\mathcal{F}$ and *probability measure* $\mathbb{P}$. The triple $\{\Omega, \mathcal{F}, \mathbb{P}\}$ is called probability space.

2.1 Sample space

A set Ω consisting of all possible outcomes (elementary events) is called sample space. The element of Ω is denoted as ω_i (for discrete outcomes). In statistical mechanics, microscopic states (microstates) are an example of elementary events which we discuss throughout this course.

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

Example: The sample space of rolling a die twice is
$\Omega = \{11, 12, 13, 14, 15, 16, 21, 22, \cdots, 65, 66\}$. There are 6^2 elements.

Example: The position of a particle $x \in \mathbb{R}$ is an sample space.

2.2 Event space

A collection of subsets of a sample space including an empty set is called an event space (we use symbol $\mathcal{F}$ for the event space). Each element in the event space is an event. Elementary events in the sample space are included in the event space. The event space for a single die is

$$
\begin{aligned}
&\emptyset, \\
&\omega_1, \cdots, \omega_6, \\
&\{\omega_1, \omega_2\}, \cdots, \{\omega_5, \omega_6\}, \\
&\{\omega_1, \omega_2, \omega_3\}, \cdots, \{\omega_4, , \omega_5, \omega_6\}, \\
&\{\omega_1, \omega_2, \omega_3, \omega_4\}, \cdots, \{\omega_3, \omega_4, \omega_5, \omega_6\}, \\
&\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}, \cdots, \{\omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}, \\
&\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}
\end{aligned}
$$

We need the event space to calculate the probabilities.

Example: If we are interested in only even number from rolling a die once, the event is $\{2, 4, 6\}$ which is a subset of sample space Ω. Thus $\{2, 4, 6\}$ is an element of the event space $\mathcal{F}$

Example: For the position of a particle, a domain $a < x < b$ is a subset of Ω. Then, the domain is an element of $\mathcal{F}$.

Event space $\mathcal{F}$ is a set consisting of a collection of subsets of Ω with the following properties:

$$\begin{array}{ll} \text{(a)} & \emptyset \in \mathcal{F} \\ \text{(b)} & \text{If } A \in \mathcal{F}, \text{ then } (\Omega \backslash A) \in \mathcal{F} \\ \text{(c)} & \text{If } A_1, A_2, \cdots \in \mathcal{F}, \text{ then } \cup_{i=1} A_i \in \mathcal{F} \end{array}$$

The event space is also called σ-algebra or σ-field because of these properties.

Example: Realizing an even number out of six possible outcomes is an event as mentioned above. Then, getting an odd number is also an event. Mathematically, if $\{2, 4, 6\} \in \mathcal{F}$ then $\{1, 3, 5\} \in \mathcal{F}$. See the property (b).

Example: Finding a particle inside a domain $a < x < b$ is an event. Then, finding the particle outside the domain is also an event. This is also property (c).

2.3 Probability measure (axioms of probability)

Probability is a map from $\mathcal{F}$ to $\mathbb{R}$ such that

$$\begin{array}{ll} \text{(a)} & P(A) \geq 0, \forall A \in \mathcal{F} \\ \text{(b)} & P(\Omega) = 1 \\ \text{(c)} & \text{For disjoint events } A_1, A_2, \cdots, P(\cup_i A_i) = \sum_i P(A_i) \end{array}$$

Example: When a die is rolled once, the corresponding sample space is $\Omega = \omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6$. We assume that $p_i = P(\omega_i) = \frac{1}{6}$ (*a priori* probability). Event of getting even number is a subset $A = \omega_2 \cup \omega_4 \cup \omega_6 \in \mathcal{F}$. The probability to obtain an even number is $P(A) = P(\omega_3 \cup \omega_4 \cup \omega_6) = P(\omega_2) + P(\omega_4) + P(\omega_6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$, which is consistent with condition (c). The probability to get any number is given by $P(\Omega) = P(\cup_i \omega_i) = \sum_i P(\omega_i) = \sum_i \frac{1}{6} = 1$, which satisfies condition (b).

3 *A priori* probability and maximum entropy principle

3.1 Equal a priori probability

The axioms of probability won't give us a way to determine the probabilities. They give us only the rules to calculate probabilities of general events from the fundamental probabilities assigned to the elemental events in Ω. So, we need to assign a probability p_i to each element ω_i in the sample space Ω. Once we know the probabilities for all elementary events, we can calculate the probabilities for all elements in the corresponding event space $\mathcal{F}$ using axiom (c). For continuous sample spaces, we consider a probability of event at the neighborhood of x as $P(\mathrm{d}x) = \rho(x)\,\mathrm{d}x$, which can be interpreted as, for example, the probability of finding a particle between x and $x+\mathrm{d}x$. The function $\rho(x)$ is called *probability density function* (PDF) or just *probability density.*

Determining the probabilities by experiments is a bit challenging since we need to repeat the experiment infinite times (cf. the Borel's law of large numbers). Since no axiom can provide the probability for elemental events in the sample space, we just have to assume it. It is called *a priori* probability. If there is no reason that the sample space is biased, we should assume that every elemental event ω_i has the same probability. This is called *principle of indifference* and the resulting probability is known as *equal a priori probability.* Statistical mechanics is based on this assumption.

Example: The sample space of single toss of an unbiased die contains six possible elementary events. "Unbiased" means that all events have identical probability. Since $P(\Omega) = \sum_i P(\omega_i) = 1$ [axiom (b)] and $P(\omega_i) = p$ [equal *a priori* probability], we have $6p = 1$. Hence, $p = \frac{1}{6}$ for all elementary events.

Example: In a segment $[a, b] \in \mathbb{R}$, a particle is found to be any point with equal probability, i.e., $\rho(x)\,\mathrm{d}x$ is constant. Using the normalization condition [axiom (b)], $P(\Omega) = \int_a^b \rho(x)\,\mathrm{d}x = \rho \int_a^b \mathrm{d}x = \rho\,(b-a) = 1$. Hence, the *a priori* probability is $\rho(x) = \frac{1}{b-a}$.

3.2 Maximum entropy principle

If there is bias among the elemental events, how can we assign a probability to each event? Obviously we cannot use the principle of indifference. We must develop a consistent mathematical method. Information entropy guides us to find the probability.

As we discussed briefly in section 1.2, information and probabilities are intimately related. The relation was originally established in information theory, it also plays a very important role in physics. Let's review it again. First we need to define information. Suppose that a coin is placed in a box. Until we open the box, we don't know which face is up. This uncertainty is the source of information. When we open the box, the uncertainty vanishes. In other words, we obtained information about the state of the coin. After opening the box, we cannot get any additional information. In information theory, it is said that 1 bit of information (see section 1.2) is acquired by opening the box and no information is left in the box. Think more complicated case, for example English language. We transmit information by writing a word. Even if the word is as short as three letters, there is a huge uncertainty. It must be one of "aaa", "aab", $\cdots$, "zzz", that is one of 26^3=17576. In terms of bit, it is $\log_2 17576 \sim 14$ bits. (This assumes that equal *a priori* probability.) This means we need 14 coins to transmit the same amount of information. When you find the word is "box", the uncertainty instantaneously vanishes and about 14 bits of information is gained. The amount of information you get is equivalent to the amount of the uncertainty. The amount of uncertainty is in turn measured by the probability, hence the amount of information also related to the probability.

In real world, the letters in the alphabet appear with different probabilities and thus equal *a priori* probability cannot be used. Shannon, who developed modern information theory, quantified the amount of information by so-called *Shannon entropy* or *information entropy*:

$$S = -\sum_i p_i \log_2 p_i$$

where p_i is the probability that event ω_i happens. When there is no uncertainty, only one of p_i is one and all others are 0. Then, $S = 1\log_2 1 + 0\log_2 0 = 0$ which is the smallest value the entropy can take (entropy is 0 or positive). What is the maximum possible value of the entropy? To find it, we try to find the maximum value of S. Noting that $\sum_i p_i = 1$, we want to find the maximum of $f(\{p_i\}) = -\sum_i p_i \log p_i - \alpha(\sum_i p_i - 1)$. using a Lagrange multiplier α:

$$\frac{\partial f}{\partial p_j} = -\log p_j - 1 - \alpha = 0$$

which implies that all p_j takes the same value. Therefore, the entropy reaches maximum value $S = -\sum_i p \log p = -\log p$ when every event carries the same probability p. If there is N elements in the sample space (N different elemental evens), $p = \frac{1}{N}$ and thus the maximum possible value of entropy is $S = \log_2 N$. For the coin box case, $N = 2$ and $p = \frac{1}{2}$. Hence, the maximum amount of information is $S = -\log_2 \frac{1}{2} = 1$. We just got the equal probability for every elemental event by maximizing the entropy without any condition. This is just the equal *a priori* probability under the principle of indifference.

Now we extend this idea for the case where there is a bias. When the bias is expressed as constraint as $g(\{p_i\}) = 0$ the probability p_i is determined by maximizing entropy under the constraint. Using another Lagrange multiplier β, we find p_i by maximizing $f(\{p_i\}) = -\sum_i p_i \log p_i - \alpha(\sum_i p_i - 1) - \beta g(\{p_i\})$. We use this method to investigate systems in canonical ensemble (Section 3.2 in the textbook.)

In statistical mechanics, we don't measure the entropy in bits. The unit of thermodynamics entropy is energy per temperature like Joule per degree. It is defined as

$$S = -k_{\mathrm{B}} \sum_i p_i \ln p_i \tag{2}$$

where k_{B} is the Boltzmann constant and natural log ln is used instead of $\log_2$.

4 Properties of probability

Here are a few basic properties of probability.

1. $P(\emptyset) = 0$.
2. $P(A) \leq 1$, $\forall A \in \mathcal{F}$.
3. $P(\Omega \backslash A) = 1 - P(A)$, $\forall A \in \mathcal{F}$.
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, $\forall A, B \in \mathcal{F}$.
5. If $A \subseteq B \subseteq \mathcal{F}$, then $P(A) \leq P(B)$.

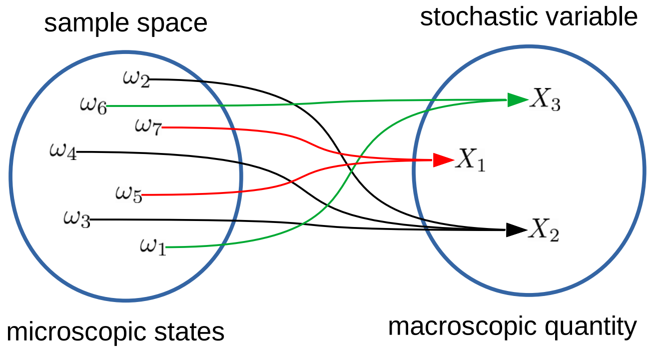

Figure 1: Stochastic variable as a map from sample space to X space. In statistical mechanics, it is a map from a microscopic states to a macroscopic quantity.

5 Stochastic variables

Let us play a gamble with a coin. I win \$1 if the outcome is head and lose \$1 otherwise. Then, we have a function $X : \Omega \mapsto \{1, -1\}$ or

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = \omega_1 = \text{"head"} \\ -1 & \text{if } \omega = \omega_2 = \text{"tail"} \end{cases} .$$

Since its value is not certain, X is called *stochastic variable*. In general, a stochastic variable is a map from sample space Ω to some quantity (See Fig. 1). Generally it is not invertible. In statistical mechanics, macroscopic quantities you measure are stochastic variables.

Example: The number of the dots on each face of a die is a stochastic variable. The sample space $\Omega = \{\omega_1 =$ ⚀, $\omega_2 =$ ⚁, $\omega_3 =$ ⚂, $\omega_4 =$ ⚃, $\omega_5 =$ ⚄, $\omega_6 =$ ⚅$\}$. Now we define a function $I(\omega) =$"number of dots on the face". Then, the value of I is determined by the outcome taken from Ω.

The probability that the stochastic variable takes a particular value, say $X = x$, is written as $P(X = x)$ or $P_X(x)$. (Lazy people like theoretical physicists often write just $P(x)$.) The uncertainty in X is obviously originated from the uncertainty in ω. Thus, if we know the probability for ω, we can find the probability for x by

$$P_X(x) = P_\Omega(\{\forall\omega; X(\omega) = x\}) .$$

Example, Let us play a gamble with a die. If the outcome of tossing a die is even, you win cash amount $\omega/2$, otherwise lose cash amount ω. So, if the die shows ⚅ you win \$3 but if it is ⚂ you lose \$3. The stochastic variable is defined as

$$G(\omega_i) = \begin{cases} i/2 & \text{if } i \text{ even} \\ -i & \text{if } i \text{ odd} \end{cases}$$

Writing the map explicitly, $\Omega = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\} \mapsto G = \{-1, 1, -3, 2, -5, 3\}$. Since this is one-to-one map,

$$P_G(g_i) = P_\Omega(\omega_i) = \frac{1}{6}$$

Now, we change the rule of the gamble. If the outcome is even, you win \$1 and otherwise lose \$1. The map is $\Omega = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\} \mapsto \{-1, 1\}$ Thus, $G = \{-1, 1\}$. This map is not invertible. $g = 1$ can be obtained from three different elementary events, ⚁, ⚃ and ⚅. Thus

$$P_G(1) = P_\Omega(\omega_2 \cup \omega_4 \cup \omega_6) = P_\Omega(\omega_2) + P_\Omega(\omega_4) + P_\Omega(\omega_6) = \frac{1}{2}$$

6 Mean

Consider a map $X : \Omega \mapsto \mathbb{R}$. The expectation value of X is defined by

$$\langle X \rangle = \sum_i X(\omega_i)\, P_\Omega(\omega_i) = \sum_j x_j\, P_X(x_j)$$

for discrete sample space $\{\omega_1, \omega_2, \cdots\}$. For continuous space,

$$\langle X \rangle = \int_\Omega X(\omega)\, \rho_\Omega(\omega)\, \mathrm{d}\omega = \int_X x\, \rho_X(x)\, \mathrm{d}x$$

where $\rho_\Omega(\omega)$ and $\rho_X(x)$ are PDF. Notice that we can calculate the mean value in the sample space Ω as well as in the space of X.

Similarly we define n-th moment

$$\langle X^n \rangle = \sum_i X^n(\omega_i)\, P_\Omega(\omega_i) = \sum_j x_j^n\, P_X(x_j)$$

or

$$\langle X^n \rangle = \int_\Omega X^n(\omega)\, \rho_\Omega(\omega)\, \mathrm{d}\omega = \int_X x^n\, \rho_X(x)\, \mathrm{d}x$$

Remark: The mean value is not necessarily the most probable value. In other words, $P(\langle \omega \rangle)$ is not necessarily the highest probability. It could be even zero, since the mean value may not happen at all. In other words, the mean value does not have to be an event.

Example: A single toss of a die has mean value $\langle I \rangle = \sum_i I(\omega_i) P_\Omega(\omega_i) = (1+2+3+4+5+6)\frac{1}{6} = 21/6 = 3.5$. The mean value is not an element of sample space. Hence the mean value does not correspond to an actual event.

The position of a Brownian particle in a harmonic potential is Gaussian distributed if it is in thermal equilibrium, meaning that its PDF of the position x is given by a Gaussian function

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-a)^2/2\sigma^2} \tag{3}$$

where a and σ are constants. The sample space is $\{x \in \mathbb{R}\}$ and the stochastic variable for the position is simply $X(x) = x$. Then, the mean position is

$$\langle x \rangle = \int_{-\infty}^{\infty} x\rho(x)\,\mathrm{d}x = a$$

which corresponds to the most probable position.

In the experiment, we often compute a mean value as

$$\bar{x}_N = \frac{1}{N}\sum_i x_i N_i$$

where N is the total number of realized events and N_i is the number of realization of event x_i. However, $\bar{x}_N$ is not a true mean value and agrees with the true mean value only when N is infinitely large, i.e., $\langle x \rangle = \lim_{N\to\infty} \bar{x}_N$, which is known as the law of large numbers. (compare this with the Borel's law of large numbers.)

7 Variance, uncertainty, fluctuations

If only one event happens without exception, i.e. $P(\omega_i) = 1$ and $P(\omega_{j\neq i}) = 0$, there is no uncertainty. Otherwise we are not sure what event will happen. Is there a quantity that measure how uncertain the outcome is? We could think of many different ways but the most popular measure of the uncertainty is variance. Suppose that stochastic variable has mean $\mu \equiv \langle X \rangle$. Then, its variance is defined by

$$\sigma^2 = \left\langle (X-\mu)^2 \right\rangle = \left\langle X^2 \right\rangle - \mu^2.$$

If there is no uncertainty, X takes only one value, that is μ. Then, the variance vanishes. The variance can diverge even when a most probable value exists. Roughly speaking the value of X lies between $\mu - \sigma$ and $\mu + \sigma$.

Example: The variance of the number obtained from the single toss of a die is

$$\sigma^2 = \left\langle I^2 \right\rangle - \langle I \rangle^2 = (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2)\frac{1}{6} - (\frac{21}{6})^2 = \frac{91}{6} - \frac{441}{36} = 2.9167$$

So, the outcome is roughly speaking between $3.5 - \sqrt{2.9} \approx 1.8$ and $3.5 + \sqrt{2.9} \approx 5.2$.

Example: If the PDE is Gaussian given by Eq (3), the variance is σ^2.

8 Joint probability and conditional probability

Suppose we have two sample spaces $\Omega = \{\omega_1, \omega_2, \cdots\}$ and $\Gamma = \{\gamma_1, \gamma_2, \cdots\}$, and th corresponding event spaces $\mathcal{F}$ and $\mathcal{G}$, respectively. The probability $P(A)$ and $P(B)$ where $A \in \mathcal{F}$ and $B \in \mathcal{G}$ are known. Let's expand the sample space to $\Omega \times \Gamma = \{\omega\gamma; \forall\omega \in \Omega, \forall\gamma \in \Gamma\}$. Then, we ask what is the probability to realize A and B simultaneously. This is known as joint probability $P(A, B)$. If the events A and B are independent, $P(A, B) = P(A)P(B)$.

When we don't know what pair of two events happen, we use the joint probability. We also encounter situations where we know event B had happened. But we want to know what is the probability that event A happens after B happened. This probability is called *conditional probability* and expresses as $P(A|B)$. Note that $P(A|B) \neq P(B|A)$ in general. The conditional probability satisfies equality

$$P(A, B) = P(A|B)P(B). \tag{4}$$

When A and B are independent, $P(A, B) = P(A)P(B)$. Then, $P(A|B) = P(A)$.

Example: We role two dice. The first die forms sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the second die $\Gamma = \{1, 2, 3, 4, 5, 6\}$. The joint sample space is $\Omega \times \Gamma = \{\{1, 1\}, \{1, 2\}, \cdots, \{6, 5\}, \{6, 6\}\}$. Since the two dice are not correlated, their outcomes are independent. Thus $P(\omega, \gamma) = P(\omega)P(\gamma) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. The conditional probability $P(\gamma|\omega) = P(\gamma) = \frac{1}{6}$ since two events are independent.

Example: There are 6 numbered balls in a jar. We take two balls out of it at random. What is the probability to obtain 1 and 4? The first ball can be any of 1 through 6. The second ball can be also 1 through 6. However, the outcomes of this example are different from the previous example because the second number is not independent from the first one since you cannot get the same number twice. Let Ω and Γ are the sample space of the first and second balls. They are the same as the previous example. However, the second ball are selected from the remaining 5 balls which depends on the first ball. Among 36 elements in $\Omega \times \Gamma$, 6 elements $\{1, 1\}, \{2, 2\}, \cdots, \{6, 6\}$ cannot be realized. Thus $P(i, i) = 0$. The remaining elements have the same probability and thus $P(1, 4) = \frac{1}{30}$. Similarly $P(4, 1) = \frac{1}{30}$. Since $\{1, 4\}$ and $\{4, 1\}$ are disjoint, the probability to get ball 1 and 4 is $P(1, 4) + P(4, 1) = \frac{1}{15}$. The conditional probability $P(1|4) = \frac{P(1,4)}{P(4)} = \frac{1/30}{1/6} = \frac{1}{5}$. The meaning of this probability is that once the first ball is known to be 4, the probability to get 1 out of the remaining five balls is $\frac{1}{5}$.

9 Correlation

Consider two stochastic variables X and Y on the same sample space Ω. If the outcome of X is statistically related to the outcome of Y, we say that they are correlated.

Example: Consider a die again. We define X and Y as follow:

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = \text{even} \\ -1 & \text{if } \omega = \text{odd} \end{cases}, \qquad Y(\omega) = \begin{cases} 1 & \text{if } \omega \leq 3 \\ -1 & \text{if } \omega \geq 4 \end{cases}$$

If the outcome of X is 1, the chance that the outcome of Y is -1 is higher since more even numbers in $\omega \geq 4$ than $\omega \leq 3$.

How can we quantify the degree of correlation? The following mean is commonly used for that:

$$C(X, Y) = \langle XY \rangle - \langle X \rangle \langle Y \rangle$$

We call $C(X, Y)$ correlation function between X and Y. It can be evaluated on the Ω as

$$C(X, Y) = \sum_i X(\omega_i)Y(\omega_i)P_\Omega(\omega_i) - \left(\sum_i X(\omega_i)P_\Omega(\omega_i) \right) \left(\sum_j Y(\omega_j)P_\Omega(\omega_j) \right)$$

or using the joint probability

$$C(X,Y) = \sum_x \sum_y xy\, P_{XY}(x,y) - \left(\sum_x x\, P_X(x) \right) \left(\sum_Y y\, P_Y(y) \right)$$

In the latter expression, if X and Y are independent $P_{XY}(x,y) = P_X(x)P_Y(y)$. Then, the correlation function vanishes, i.e. $C(X,Y) = 0$ if X and Y are not correlated.

Example: For the previous example, the mean of each stochastic variables vanish and thus $\langle X \rangle = \langle Y \rangle = 0$.

$$\begin{aligned} C(X,Y) &= \sum_i X(\omega_i) Y(\omega_i) P_\Omega(\omega_i) \\ &= [(-1)\cdot 1 + 1 \cdot 1 + (-1)\cdot 1 + 1\cdot(-1) + (-1)\cdot(-1) + 1\cdot(-1)]\,\frac{1}{6} \\ &= -\frac{1}{3} \end{aligned}$$

10 Typicality: an example

In this section, we illuminates the basic idea of statistical mechanics using a simple model. Consider N distinguishable balls. Here "distinguishable" means that each ball has a name on it. We use name b_i for the i-th ball and we shall call the set of balls *system*. The balls are placed in one of two urns labeled with "0" and "1", respectively. The *state* of each ball s_i is specified by "0" or "1", i.e. if the i-th ball is in urn 0, then $s_i = 0$ and in the other urn $s_i = 1$. The *microscopic state* or *microstate* of the system is denoted as $\omega = s_1 s_2 \cdots s_{\mathrm{N}}$, which is a string of "0" and "1". For example, if 3 balls are in urn 0 and 1 ball in the other urn, the state is `0001`. We assume that there is no interaction between the balls. Such a system is called *ideal gas*.

The microscopic state of classical particles or electron spins changes as time goes based on the Newton's laws of motion or the Shchrödinger equation. In the current model, we assume a simple rule of dynamics. We pick a ball at random and move it from the current urn to the other. In other words, the state of the ball flips between 0 and 1. Accordingly, the microscopic state of the system also changes but only one letter in the string changes at one time, like `0010` $\to$ `1010`.

Now we consider a macroscopic state. Our eyes are not good enough to see individual balls. We can see only the total number of balls in each urn. Since the total number of balls conserves, we need to investigate only the number of balls in urn 1. Writing it in a mathematical form, we define a stochastic value $N_1(\omega) = \sum_i s_i$. This event is a macroscopic quantity we want to measure. Since the state of the system changes as time goes ω is a function of time and so is N_1. Any dynamics must begin with an initial state. We hope that the longtime behavior of N_1 does not depend on the initial condition. The present statistical mechanics usually assume that macroscopic states will approach a steady state or more precisely *thermodynamics equilibrium*. However, that is not always guaranteed.

Before solving this problem, we summarize the mathematical structure of the problem based on the probability theory. First, we need a sample space. In the statistical mechanics, the sample space is a set consisting of all possible microscopic state. In the current model system, the strings

Table 1: Sample and event spaces for the system with 4 balls

Sample space		Events		
sample ID	microstates	event ID	# of balls in 0	size
ω_1	0000	e_0	0	1
ω_2	0001			
ω_3	0010	e_1	1	4
ω_4	0100			
ω_5	1000			
ω_6	0011			
ω_7	0101			
ω_8	1001	e_2	2	6
ω_9	0110			
ω_{10}	1010			
ω_{11}	1100			
ω_{12}	0111			
ω_{13}	1011	e_3	3	4
ω_{14}	1101			
ω_{15}	1110			
ω_{16}	1111	e_4	4	1

of 0 and 1 are the element of the sample space. For example, if we have 4 balls, there are $2^4 = 16$ different microscopic states listed in Table 1.

Microscopically, the system changes from ω_i to ω_j and we are interested in how the change in the miscrostate results in the change in the macroscopic state. Formally, we should investigate the time evolution of event. However, it is often difficult to study the dynamics. On the other hand, our experience suggests that the macroscopic quantity won't change much within the resolution of our eyes after a certain period of time. Is there are any way to find the steady state value without directly solving the dynamics? To do so, we need to look at the rule of dynamics. Random choice of a single ball is not affected by the state of other balls. That means the dynamics does not depend on the value of the current state. It jumps from one misrostate to another without any bias, which implies that every microstate happens with an equal probability. This "assumption" sets the *a priori* probability. Since there is 2^N microstates, the *a priori* probability for each microstate is $p_i = \dfrac{1}{2^N}$.

Next we find the probability that each event take place. Note there are $N+1$ events. We denote event e_{N_1} where N_1 is the number of balls in urn 1. The number of particles in urn 0 is $N_0 = N - N_1$. e_0 means all balls are in urn 0 and the other urn is empty. There is only microstate for e_0. It is easy to find that event e_1 consists of N microstates (see Table 1 for four-ball system.) In general case we need to count the number of string which has N_1 of 1 and N_0 of 0. This is an elementary combinatorial calculation and the size of e_{N_1} is simply a binomial coefficient $\dbinom{N}{N_1} = \dfrac{N!}{N_1!N_0!}$. Using the axiom (c) of probability,

$$P(e_{N_1}) = \sum_{\omega \in e_{N_1}} P(\omega) = \frac{N!}{N_1!N_0!}\frac{1}{2^N} \tag{5}$$

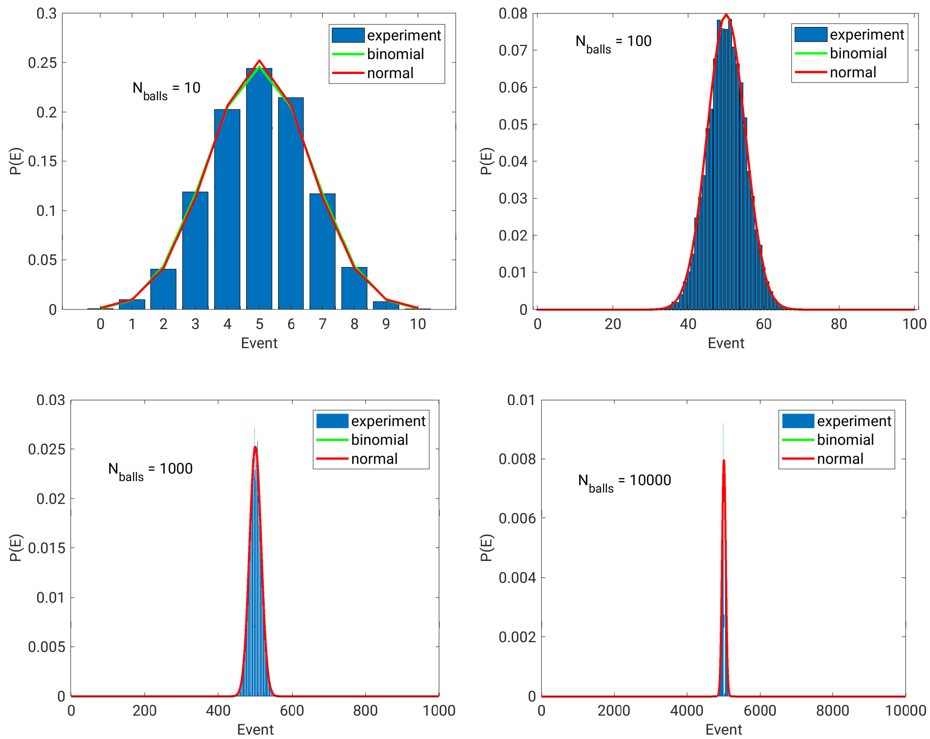

Figure 2: Binomial distribution with $p = 0.5$. As the number of trials increases, clearly $k \approx N/2$ becomes dominant.

which is a special case of binomial distribution.

Here is a brief summary of binomial distribution. Suppose that you pick either 0 or 1 with probability q and $1 - q$, respectively. Suppose that out of N particle, N_0 particles go to urn 0 and N_1 go to urn 1. What is the probability of that event? simple combinatrial counting tells that

$$P(N_1; N) = \frac{N!}{N_0!N_1!} q^{N_0}(1-q)^{N_1}. \tag{6}$$

Equation (5) is a special case of Eq. (6) where $q = \frac{1}{2}$. The mean and variance of the binomical distribution are $\mu = \langle N_1 \rangle = Nq = \frac{N}{2}$ and $\sigma^2 = \frac{N}{4}$. For $N \gg 1$, the binomial distribution is very close to the normal distribution with the same mean and variance, that is

$$p(N_1; N) \quad \xrightarrow{N\to\infty} \quad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(N_1-\mu)^2/2\sigma^2} \tag{7}$$

In figure 2, the binomial distribution is plotted for several different numbers of balls. When $N = 10$, The probability that all balls are in one urn is still visible although the chance is very small. Eight balls in 0 and two in 1 will happen with a significant probability. However, as the number of ball increases, $N_1 \approx N/2$ dominates and other possibilities nearly vanish. Despite that every microstate is realized with an equal probability, we see that the balls are almost evenly split into two urns. We say that $N_1 = N/2$ is a typical state. In the limit of $N \to \infty$, the uncertainty

Table 2: Ratio $P(k = 0.5N)/P(k = 0.4N)$ for various N. As N increases, the ratio grows exponentially. For $N > 1000$, the chance that 40% of the particles in urn 0 virtually vanishes.

# of balls	$P(k = 0.5N)/P(k = 0.4N)$
10	1.2
100	7.3
1000	5.4×10^{8}
10000	2.7×10^{87}
100000	2.9×10^{874}

vanishes and the typical event (50%-50% split) is guaranteed from the macroscopic view. This limit is called *thermodynamic limit*. Statistical mechanics tries to find this typical value.

In terms dynamical picture, the system moves from one microstate to another without bias. It visits every state eventually. However, macroscopic value e won't change because almost all microstates have the same macroscopic values. Mathematically writing it, $P_\Omega(\omega_1) = P_\Omega(\omega_2) = \cdots = P_\Omega(\omega_N)$ but $P(N_1 = \frac{N}{2}) \approx 1$. To demonstrate it, the ratio $P(k = 0.5 * N)/P(k = 0.4 * N)$ for different N are shown in Table 2. When $N = 10$, $N_1 = 0.5N$ and $N_1 = 0.4N$ are equally probable. However, as N increases the ratio grows exponentially. For $N > 1000$, $P(N_1 = 0.4N)$ is vanishing small. Can you imagine the ratio at $N = 10^{23}$? You can see only 50%-50% split without exception during your life time even when microstates are changing rapidly.

In summary, for large systems, a macroscopic quantity takes a typical value despite that all microstates happen with the equal probability. The probability for any atypical event is vanishingly small and we are not able to se the event during a limited observation time. Now, how can we find a typical value? It is simple because the mean value is exactly the typical value because the variance vanishes in the thermodynamic limit. See Fig. 2 shows that the disitrbution is very narrow for $N = 10000$.

Remark 1: The above procedure involves some controversial procedure. We replace the time evolution with ensemble. We are supposed to take snapshot of the system many time as time evolves. The eqch snapshot is a microstate at the moment. From the snapshot, we evaluate the macroscopic quantity and take statistical over many snapshots. Instead, we prepare many identical systems. We take a single snapshot of each system in the ensemble and take statistical analysis. We assume that the outcome of these two statistical analysis agree each other. However, that is not guaranteed in real world problems. This is the issue of Ergodic hypothesis. We hoped that non-linear chaotic dynamics resolves this issue but no one is er able to prove it. The resolution seems coming out of unexpected area. Quantum entanglement in a large quantum system is sufficient to justify the use of ensemble approach. This is currently a hot research topics.

Remark 2: The above In the above analysis, we assume that the balls are not interacting to each other. In real world, particles collides through both attractive and repulsive coupling. The interaction bring many interesting thermodynamic behavior such as critical phenomena. When the coupling is small, we can use perturbation theory but such approximation is often useless to study critical phenomena. The theoretical investigation is quite challenging when the interaction is strong.

Remark 3: We also assumed that the balls are distinguishable. In the quantum world, the same kind of particles are not distinguishable. Then, we have to use a different counting rules, or precisely different statistic, namely Bose-Einstein and Fermi-Dirac statistics. In low dimensional (1d and 2d) world the story is even more complicated. However, quantum statistics is a major source of many intersting phenomena such as superconductors.

Remark 4: Probability theory is closely related to information theory. In the quantum world information behave differently, which is the source of quantum computers and other quantum information processess. Statistical mechanics is deeply dependent on the quantum information due to quantum entanglement. We are working hard to understand it.

11 Central limit theorem

Suppose that stochastic variable X has mean value μ and variance σ^2. Let $X_1, X_2, \cdots, X_n$ n realization of X. Then, we define a new stochastic variable $S = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$. The central limit theorem states that as $n \to \infty$, the probability distribution of S approaches a normal distribution

$$P_S(s) \to \frac{1}{2\pi\sigma_n^2} e^{-\frac{(s-\mu)^2}{2\sigma_n^2}} \tag{8}$$

where $\sigma_n^2 = \frac{\sigma^2}{n}$. This means that the mean value of S is the same as that of X. This is nothing but the law of large number. On the other hand, the variance of S is smaller than that of X by factor $\frac{1}{n}$. For large n, the uncertainty of S vanishes and S take only one value μ. This situation is similar to the above discussion on the typicality if X is the microscopic space and S is a macroscopic quantity.

12 Appendix

12.1 Evaluation of binomical coefficients with Mathematica

Let's compute the number of microstates for case 1: 50% in 0 and 50% in 1 and case 2: 40% in 0 and 60% in 1. Then, we computer the ratio. Using the binomial coefficient $\begin{pmatrix} N \\ M \end{pmatrix}$ we compute the following with Mathematica.

$$\begin{pmatrix} K \\ K*0.5 \end{pmatrix} / \begin{pmatrix} K \\ K*0.4 \end{pmatrix} \text{ where K = number of balls}$$

In: **R = Table[N[Binomial[K, K * 0.5]/Binomial[K, K * 0.4]],**

{K, {10, 100, 1000, 10000, 100000}}];

In: **Do[Print["K=", 10^K, " Binomial=", R[[K]]], {K, {1, 2, 3, 4, 5}}]**

K=10 Binomial=1.2
K=100 Binomial=7.33956
K=1000 Binomial=5.44357×10^8
K=10000 Binomial=2.74511×10^{87}
K=100000 Binomial=$2.9199688066 \times 10^{874}$

12.2 Monte Carlo simulation with Python

Python source code:

```
#!/usr/bin/env python3

# import packages
import numpy as np                # package for numerical calculation
import scipy.stats as ss          # package for statistical calculation
import matplotlib.pyplot as plt   # package for plotting

# set parameter values
n_balls=10          # number of balls
n_trials=10000      # number of trials in Monte Carlo simulation

# set array size
h=np.zeros(n_balls+1)
r=np.zeros(n_balls)

# Monte Carlo simulation
for n in range(1,n_trials):
r=np.random.rand(n_balls)  # generate a random numbver for each ball
n=sum(r>0.5)               # count the number of balls in urn 1
h[n]=h[n]+1                # record the result in histgram

# normalize the distribution
h=h/n_trials

# event values
x=np.linspace(0,n_balls,n_balls+1)

# binomial distribution (exact theory)
y_binomial=ss.binom.pmf(x,n_balls,0.5)

# normal distribution (approximation to binomial pdf)
y_normal=ss.norm.pdf(x, 0.5*n_balls,np.sqrt(n_balls*0.25))

# plot the data
plt.bar(x, h, align='center',label='experment')
plt.plot(x,y_binomial,label='binomial',color="green")
plt.plot(x,y_normal,label='normal',color="red")
plt.xlabel('Event')
plt.ylabel('P(E)')
plt.legend(loc=1)
plt.text(n_balls*0.1,y_binomial[int(n_balls*0.5)]*0.9,r'$N_{ball}=$'+str(n_balls))
plt.show()
```

12.3 Monte Carlo simulation with Matlab

MATLAB source code:

```
% Set parameter values
n_balls=10000;    % number of balls
n_trials=10000;   % number of trials in Monte Carlo simulation

% Set array size
h=zeros(1,n_balls+1);
r=zeros(1,n_balls);

% Monte Carlo simulation
for i=1:n_trials
r=rand(1,n_balls);  % generate randum numbers for each ball
n=sum(r>0.5);       % count the number of balls in urn 1
h(n+1)=h(n+1)+1;    % record the result in histgram
end

% normalize the distribution
h=h/n_trials;

% event values
x=linspace(0,n_balls,n_balls+1);

% binomial distribution (exact theory)
y_binomial=binopdf(x,n_balls,0.5);

% normal distribution (approximation to binomial pdf)
y_normal=normpdf(x,n_balls*0.5,sqrt(n_balls*0.25));

% plot the data
q=bar(x,h,'DisplayName','experiment');
hold on

p1=plot(x,y_binomial,'DisplayName','binomial');
set(p1,'linewidth',2,'color','green');

p2=plot(x,y_normal,'DisplayName','normal');
set(p2,'linewidth',2,'color','red');

legend('show');

xlabel('Event')
ylabel(texlabel('P(E)'));
```

```
txt=['N_{balls} = ',num2str(n_balls)];
text(n_balls*0.1,y_binomial(n_balls/2+1)*0.9,txt);
hold off
```

13 Homework

Read the following investigation and answer to questions Q1, Q2, Q3, and Q4.

When we rolled a die six times, outcome { ⚂, ⚄, ⚀, ⚁, ⚀, ⚅ } hardly surprises us. On the other hand, we feel very lucky when { ⚀, ⚀, ⚀, ⚀, ⚀, ⚀ } happens. However, these two outcomes happen with the same probability. So, why are we surprised by all ⚀? When we see the first outcome, we note that ⚀ is realized twice and others are all different. Our brain forgets the details after a while but still remember that the same face appeared twice. In other words, we don't look at details and capture only a few simple features. Since having same faces twice in six tries is very ubiquitous, we are not surprised at all. We will investigate it quantitatively.

1. Sample space: Microstates

There are N elementary events:

$$\begin{aligned} \omega_1 &= \text{⚀⚀⚀⚀⚀⚀} \\ \omega_2 &= \text{⚀⚀⚀⚀⚀⚁} \\ &\vdots \\ \omega_N &= \text{⚅⚅⚅⚅⚅⚅} \end{aligned}$$

we shall call these events microstates.

Q1: Find N,

2. *A priori* probability

The die is assumed to be unbiased and the maximum entropy principle should work.

Q2: Find the probability that each miscrostate is realized.

3. Stochastic variable

Let us assume that our brain and eyes are not sharp enough to recognize individual microstates. We see only faces that appear more than twice. The macroscopic quantity M we observe is the largest number of the repeated faces in the realizations. M is a stochastic variable. In Table 3, the map from microscopic states to macroscopic is listed. Subsets of miscrostates are represented with a string like "aabcde" which means two outcomes are the same but the remaining four are different. Its permutation is included in the same subset.

You can see that aabcde is 1800 more likely than aaaaaa. That is why we think { ⚀, ⚀, ⚀, ⚀, ⚀, ⚀ } is rare but { ⚂, ⚄, ⚀, ⚁, ⚀, ⚅ } is ubiquitous despite that both have the same probability.

Q3: Verify all numbers in Table 3.

Q4: Are you surprised if the outcome is { ⚁, ⚄, ⚀, ⚀, ⚃, ⚄ }? Note that ⚀ and ⚄ appear twice. Explain what you feel based on statistics.

Table 3: Macroscopically ubiquitous events

Macroscopic events	Microstates	Number of microstates	Probability
M_1	abcdef	720	0.015
M_2	aabcde	10800	0.231
	aabbcd	16200	0.347
	aabbcc	1800	0.039
M_3	aaabcd	7200	0.154
	aaabbc	7200	0.154
	aaabbb	300	0.006
M_4	aaaabc	1800	0.039
	aaaabb	450	0.010
M_5	aaaaab	180	0.004
M_6	aaaaaa	6	0.0001
Total	6^6	46656	1