A Note on Probability Theory

Ying Nian Wu, Note for STATS 200A

Contents

1 Probability

1.1 Why probability?

According to Maxwell, "the true logic of this world is in the calculus of probabilities."

The most fundamental physics laws in quantum mechanics are probabilistic, where the wave function $\phi(x,t)$ of a system evolves according to the Schrodinger equation, and $p(x,t) = |\phi(x,t)|^2$ tells us the probability of finding the system at state x at time t.

For statistical physics that studies systems with large numbers of elements, the population or ensemble of the states that a system can take is described by a probability distribution. The phenomenon of phase transition can be explained by the fact that the probability distribution of a system can change if the boundary condition changes.

The field of machine learning is about learning from training examples, and generalizing to the testing examples. Both training examples and testing examples can be considered random samples from a population or probability distribution. Learning is about estimating properties of the probability distribution based on a finite number of training examples.

The Monte Carlo method uses random sampling for computation. The error of a Monte Carlo method does not depend on the dimension of the problem, so that the Monte Carlo method is often the only method that can work in the high-dimensional situation.

Signal compression, error correction coding, Google page rank, etc. are with us in our daily life. They are all based on probabilistic modeling.

1.2 Three canonical examples

We may use the following three examples to think about probability.

Example 1: Roll a fair die. The probability of getting any number in $\{1,2,...,6\}$ is 1/6. The probability of getting a number greater than 4 is 2 out of 6, which is 1/3.

Example 2: Randomly sample a person from a population. The probability of getting a male is the proportion of male sub-population. The probability of getting a person taller than 6 feet is the proportion of the sub-population of people who are taller than 6 feet.

Example 3: Randomly throw a point into a region Ω. The probability of the point falling into a sub-region A inside Ω is the area of A divided by the total area of Ω.

In each of the above three examples, the outcomes (numbers, persons, positions) are equally likely. Probability is a common sense of uncertainty.

1.3 Long run frequency

Probability also manifests itself as long run frequency. For instance, if we flip a fair coin, the probability of getting a head is 1/2. If we flip the fair coin many times independently, the frequency of heads approaches 1/2. So the probability of an event can be interpreted as how often it happens.

Can we define probability of an event as its long run frequency? The answer is no. For instance, if we flip a fair coin n times, there are 2^n possible sequences. Some sequences do not have frequencies approaching 1/2. For instance, the sequence of all heads has frequency equal to 1. The fact that the frequency approaches 1/2 is true only with high probability (approaching 1) if we assume that all the 2^n sequences are equally likely. So we should define probability first, and then quantify the meaning of frequency approaching probability as a high probability event.

1.4 Basic language and notation

We call the phenomenon under study an experiment. The experiment generates a random outcome. The set of all the outcomes is called the sample space. In this course, we use Ω to denote the sample space, and we use ω to denote an outcome. We usually work with some numerical descriptions of the outcome, denoted by, e.g., $X(\omega)$.

In Example 1, the outcome itself is a number.

In Example 2, ω is a random person, $X(\omega)$ may be the height of the person, and $Y(\omega)$ may be the gender of the person.

We call X a random variable, although strictly speaking, X is a mapping or a function that maps each $\omega \in \Omega$ to a number.

An event is a subset of the sample space. It can be described by a statement in words about the outcome. It can also be described by a mathematical statement about the corresponding random variable. In any case, an event is the subset of all the outcomes that satisfy the statement. Usually we use the notation A, B, C to denote an event.

In Example 2, the sample space is the population, and an event is a sub-population. Let A be the event that the person is taller than 6 feet, then A is the sub-population of people taller than 6 feet. We can write

$$A = \{\omega : X(\omega) > 6\}.$$

The above expression connects event and random variable.

Probability is defined on the event. For instance,

$$P(A) = P(\{\omega : X(\omega) > 6\}) = P(X > 6).$$

We often simplify the notation $X(\omega)$ into X.

There are three relations between the events: (1) AND or intersection, $A \cap B$. (2) OR or union, $A \cup B$. (3) NOT or complement A^c or $\overline{A}$.

1.5 Axioms

Probability behaves like size or measure.

In Example 2, an event A is a sub-population, $P(A) = |A|/|\Omega|$ is the proportion of this sub-population, where $|A|$ is the number of people in A, and $|\Omega|$ is the total number of people.

In Example 3, $P(A) = |A|/|\Omega|$, where $|A|$ is the area of A, and $|\Omega|$ is the total area.

We use the following axioms to quantify our notion of probability measure.

- For any event A, $P(A) \geq 0$.
- For the sample space Ω, $P(\Omega) = 1$.
- If $A \cap B = \phi$, then $P(A \cup B) = P(A) + P(B)$.

The last property is called additivity. By induction, it implies finite additivity, i.e., if $A_1, ..., A_n$ are disjoint, i.e., $A_i \cap A_j = \phi$ for $i \neq j$, then $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$. In modern probability theory, we assume countable additivity, i.e., if $A_i, i = 1, 2, ...$ are disjoint, then $P(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$.

Other rules can be proved using the axioms: (1) $P(A^c) = 1 - P(A)$. (2) If $A \subset B$, then $P(A) \leq P(B)$. (3) Inclusion-exclusion: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

1.6 Sigma algebra

Probability is not defined for all the subsets of Ω. It is defined for subsets that belong to a σ-algebra $\mathscr{F}$, which can be considered the collection of all the meaningful statements. It satisfies the following two conditions:

(1) If A is meaningful, then A^c is also meaningful.

(2) If A_1, A_2, ... are meaningful, then $\cup_{i=1}^\infty A_i$ is also meaningful.

Because $(\cup A_i)^c = \cap A_i^c$, we also have

(3) If $A_1, A_2, ...$ are meaningful, then $\cap_{i=1}^{\infty} A_i$ is also meaningful.

That is, $\mathscr{F}$ is closed under logical compositions. It is an algebra because it is closed under complement, union and intersection. It is a σ-algebra because the union and intersection can be countable union and countable intersection.

A σ-algebra can be generated by a set of basic events or simple statements, by combining them by complement, countable union and intersection. This method will create the smallest σ-algebra that contains these simple events. For instance, if $\Omega = [0,1]$, the basic statements can be intervals $(a,b) \subset [0,1]$. The σ-algebra generated by the intervals is called a Borel σ-algebra.

As another example, if we flip a fair coin independently, Ω contains all the infinite sequences of heads and tails. The simple statements include all the statements about the outcome from the first n flips.

1.7 Why sigma-algebra

Why do we need sigma-algebra? i.e., why do we need to entertain events that are countable unions and/or intersections of simple events or statements? We shall illustrate the reason by a special case of strong law of large number.

Let ω be a sequence of heads and tails, and let $X_i(\omega)$ be the result of the i-th flip, so that $X_i(\omega) = 1$ if the i-th flip is head, and $X_i(\omega) = 0$ if the i-th flip is tail. Let $\bar{X}_n = \sum_{i=1}^n X_i(\omega)/n$ be the frequency of heads in the first n flips. Let

$$A_{n,\varepsilon} = \{\omega : |\bar{X}_n(\omega) - 1/2| < \varepsilon\}$$

collects the sequences whose frequencies are close to 1/2. The weak law of large numbers says that

$$P(A_{n,\varepsilon}) = \frac{|A_{n,\varepsilon}|}{2^n} \to 1,$$

for any fixed ε, where $|A_{n,\varepsilon}|$ denotes the number of sequences of the first n flips whose frequencies of heads are close to 1/2.

The strong law of large numbers says that $P(\bar{X}_n \to 1/2) = 1$. $\bar{X}_n \to 1/2$ means that for any $\varepsilon > 0$, there exists N, such that for any $n \geq N$, $A_{n,\varepsilon}$ is true, so the event

$$A = \{\omega : \bar{X}(\omega) \to 1/2\} = \cap_{k=1}^{\infty} \cup_{N=1}^{\infty} \cap_{n=N}^{\infty} A_{n,\varepsilon=1/k},$$

which means for any ε, $A_{n,\varepsilon}$ happen infinitely often. The strong law of large numbers is $P(A) = 1$. The strong law of large numbers illustrates the need for countable union and intersection.

2 Measure

2.1 What is measure?

As mentioned above, in the equally likely setting, $P(A) = |A|/|\Omega|$, where $|A|$ denotes the size of A, and $|\Omega|$ denotes the size of Ω. Also as mentioned above, this observation motivates us to define probability as measure.

We all have the intuitive notion of measure, such as count, length, area, volume, etc. But how do we develop a logically coherent theory of measure?

Let us assume $\Omega = [0,1]^2$. For any $A \subset \Omega$, how do we define its area? Can we always define an area for any A?

The following are the logical steps to develop a theory of measure.

(1) We start from simple shapes, such as Jordan simple shapes, each of which is a finite union of non-overlapping rectangles. The measure of each Jordan simple shape A, denoted by $\mu(A)$, is the sum of the areas of these rectangles.

(2) For each set A, we define its outer measure

$$\mu^*(A) = \inf_{A \subset S} \mu(S),$$

as the minimal cover, where S is a simple shape that covers A. inf means minimum, but it may not be attainable by any S, although it can be the limit for a sequence of S. For instance, the region inside a circle has an outer measure, but it cannot be attained by Jordan simple shapes. However, it can be approached by Jordan simple shapes.

(3) For each set A, we define its inner measure

$$\mu_*(A) = \sup_{S \subset A} \mu(S),$$

as the maximal filling, where S is a simple shape that is inside A. sup means maximum. Again it may only be approached as a limit.

(4) A is measurable is $\mu^*(A) = \mu_*(A)$, and we denote it as $\mu(A)$. We can collect all the measurable sets, and define measure for this collection.

2.2 Lebesgue measure

For Lebesgue measure, the simple shapes are countable union of non-overlapping rectangles. But there is a subtlety. The simple shapes are used to define outer measure, i.e., the simple shapes are the covering shapes. To define inner measure, we need filling shapes, and they are like the complements of the covering shapes. This is due to the fact that we can define the inner measure as

$$\mu_*(A) = 1 - \mu^*(A^c),$$

where the complement of the cover of A^c is the filling of A.

For Jordan simple shapes, their complements are still Jordan simple shapes. But we need to be more careful with Lebesgue simple shapes.

The above complement definition also motivates a splitting definition, that is A is measurable if for any B,

$$\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c).$$

The splitting condition is convenient for proving things, although it is less intuitive.

The Lebesque measurable sets form a σ-algebra. But the Jordan measurable sets do not form a σ-algebra.

2.3 Law of large number

Equipped with the Lebesgue measure, let us revisit the law of large number, by studying a more geometric case.

Let $X_1, ..., X_n \sim \text{Unif}[0,1]$ independently, the law of large number says that

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \to 1/2,$$

in some sense as $n \to \infty$.

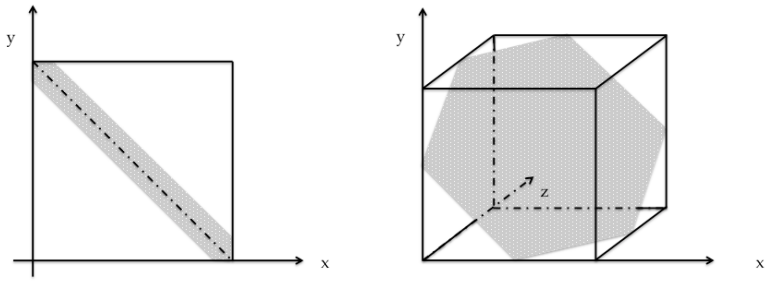

Figure 1: Geometry of the weak law of large number. The central diagonal piece of thickness ε occupies most of the volume for large n, no matter how small ε is. Here we show $n = 2$ and $n = 3$.

The weak law of large number is

$$P(|\bar{X}_n - 1/2| < \varepsilon) \to 1,$$

for any $\varepsilon > 0$. This is called convergence in probability.

The strong law of large number is

$$P(\bar{X}_n \to 1/2) = 1,$$

which is called almost sure convergence.

There is a geometric meaning of the weak law of large numbers. It means that for the n-dimensional cube $[0,1]^n$ whose volume is 1, the volume of the diagonal piece,

$$A_{n,\varepsilon} = \{(x_1,...,x_n) \in [0,1]^n : \frac{1}{n}\sum_{i=1}^{n} x_i \in [1/2-\varepsilon, 1/2+\varepsilon]\},$$

will approach 1, i.e., the central diagonal piece almost occupies the volume of the whole cube, as $n \to \infty$.

The strong law of large number is about infinite sequence $X_1, X_2, ... \sim \text{Unif}[0,1]$ independently. The sample space is the infinite dimensional cube $[0,1]^\infty$. Let

$$A = \{(x_1, x_2, ...) \in [0,1]^\infty : \frac{1}{n}\sum_{i=1}^{n} x_i \to 1/2\}$$

be the set of convergent sequences, where each sequence is a point in the infinite dimensional cube. Again

$$A = \cap_{k=1}^{\infty} \cup_{N=1}^{\infty} \cap_{n=N}^{\infty} A_{n,\varepsilon=1/k}.$$

We may think of A as the arbitrarily thin diagonal piece of the infinite dimensional cube. The strong law of large number says that the volume of A is 1.

2.4 Concentration of measure

The law of large number results from a phenomenon called concentration of measure. According to the Hoeffding inequality, which is a concentration inequality,

$$\mu(A_{n,\varepsilon}^c) \leq 2e^{-2n\varepsilon^2},$$

which is a sharp bound of the off-diagonal pieces or the tails. This directly implies the weak law of large number.

As to the strong law of large number,

$$A^c = \cup_{k=1}^{\infty} \cap_{N=1}^{\infty} \cup_{n=N}^{\infty} A_{n,\varepsilon=1/k}^c.$$

For fixed ε, we have

$$\mu(\cap_{N=1}^{\infty} \cup_{n=N}^{\infty} A_{n,\varepsilon}^c) \leq \mu(\cup_{n=N}^{\infty} A_{n,\varepsilon}^c) \leq \sum_{n=N}^{\infty} \mu(A_{n,\varepsilon}^c)$$

for any fixed N. The left hand side goes to 0 as $N \to \infty$, because

$$\sum_{n=1}^{\infty} \mu(A_{n,\varepsilon}^c) \leq \sum_{n=1}^{\infty} 2e^{-2n\varepsilon^2} < \infty.$$

Thus $\mu(A^c) = 0$ and $\mu(A) = 1$. The above inequality is the condition of the Borel-Cantelli lemma, which is a tool for proving the strong law or almost sure convergence.

One caveat is about the definition of $A_{n,\varepsilon}$, which we were not careful or precise in the above derivation. For the weak law, $A_{n,\varepsilon} \in \Omega_n = [0,1]^n$, the n-dimensional cube. But for the strong law, $A_{n,\varepsilon} \in \Omega = [0,1]^{\infty}$, i.e., the infinite dimensional cube,

$$A_{n,\varepsilon} = \{(x_1, x_2, ...,) \in [0,1]^{\infty} : \frac{1}{n}\sum_{i=1}^{n} x_i \in [1/2 - \varepsilon, 1/2 + \varepsilon]\}.$$

If we want to be more careful, we may denote it as $\tilde{A}_{n,\varepsilon}$. Compared to the $A_{n,\varepsilon}$ in the weak law,

$$\tilde{A}_{n,\varepsilon} = A_{n,\varepsilon} \times [0,1] \times [0,1] \times ...$$

and

$$\mu(\tilde{A}_{n,\varepsilon}) = \mu(A_{n,\varepsilon}) \times 1 \times 1 \times ... = \mu(A_{n,\varepsilon}).$$

So the above proof goes through. But we must be clear that $\tilde{A}_{n,\varepsilon}$ is a set in the infinite dimensional cube. Because of the infinite additivity of the Lebesgue measure, we can measure A, and prove the strong law via the Borel-Cantelli lemma. For the weak law to hold, we only need Jordan measurability.

The above proof of the strong law of large number can be applied to the coin flipping case, based on the same concentration inequality. Here $P(A)$ or $\mu(A)$ is $|A_{n,\varepsilon}|/2^n$, i.e., the proportion of sequences whose frequencies of heads are close to 1/2 among all the 2^n sequences, where $|A_{n,\varepsilon}|$ counts the number of sequences in $A_{n,\varepsilon}$. Geometrically, each sequence is a vertex in the n-dimensional cube, and $A_{n,\varepsilon}$ collects the vertices within the diagonal piece. But this geometric intuition is not very helpful.

2.5 Lebesgue integral

For a continuous random variable $X \sim f(x)$, we have $P(X \in A) = \int_A f(x)dx$, and $\mathbb{E}(h(X)) = \int h(x)f(x)dx$. Both are integrals. In measure theoretical probability, they are Lebesgue integrals.

The Lebesgue integral is a re-tooling of the Riemann integral. They differ in how to discretize. The Riemann integral discretizes the domain of the function, whereas the Lebesgue integral discretizes the range of the function. The following figure from Wikipedia provides a good illustration. Why do we need to

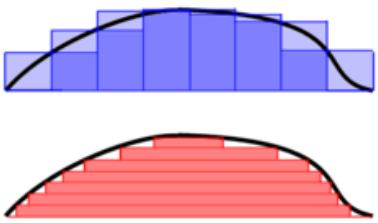

Figure 2: Illustrations of the Riemann integral (above) and the Lebesgue integral (below). Source: Wikipedia

discretize in the Lebesgue's way? The main reason is that it can handle the limit more conveniently. Specifically, if we have a sequence of functions $f_k(x) \to f(x)$ for each x in the domain, then with very mild conditions, $\int f_k \to \int f$ for Lebesgue integral. This can be very useful in Fourier analysis and functional analysis, but this property is sometimes not enjoyed by the Riemann integral. The reason is actually very easy to understand. Since the limit $f_k(x) \to f(x)$ takes place in the range of f with x fixed, it is more natural to discretize the range so that the limit can be more easily handled.

For the Lebesgue integral of a positive function $f > 0$, suppose we discretize the range into equally spaced horizontal bins, $[y_i, y_{i+1} = y_i + \Delta y)$. Then according to the above figure, we can define

$$\int f = \lim_{\Delta y \to 0} \sum_i \mu(\{x : f(x) > y_i\}) \Delta y, \tag{1}$$

where $\mu(\{x : f(x) > y_i\})$ is the length or the Lebesgue measure of the horizontal line segment in which the function f is above y_i. $\mu(\{x : f(x) > y_i\})\Delta y$ is the area of the horizontal bin immediately beneath y_i. We can denote $\int f$ by $\int f d\mu$ or $\int f(x)\mu(dx)$.

So we must be able to measure the length of $\{x : f(x) > y\}$ for every y. Such a function f is called a measurable function. Roughly speaking, it means that $\{x : f(x) > y\}$ can be broken up into countably many disjoint intervals, so that the total length of $\{x : f(x) > y\}$ can be measured as the sum of the lengths of these disjoint intervals according to the infinite additivity. Or in other words, $\{x : f(x) > y\})$ belongs to $\mathscr{L}$.

For a more general f that is not always positive, we can separately integrate its positive part and negative part, and then calculate their difference.

As an intuitive example, we can imagine a cashier collecting money. He can sum up the bills over time as he collects them. But he can also place the bills into different drawers. In the end, he can calculate the amount of money in each drawer, and then add them up. Lebesgue himself used such an example although he did not invoke the analogy of "cashier" and "drawers".

2.6 Simple functions

The above definition of the Lebesgue integral is constructive. A less constructive but more elegant definition is to consider simple functions of the form $s(x) = \sum_i a_i 1_{S_i}(x)$, where $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ otherwise. So $s(x)$ is a piecewise constant function. We require that each S_i is measurable, i.e., its length $\mu(S_i)$ is defined, i.e., $S_i \in \mathscr{L}$. We do not require that S_i are disjoint. We can then define

$$\int s(x)\mu(dx) = \sum_i a_i \mu(S_i).$$

Then we can define

$$\int f = \sup_{s \le f} \int s.$$

In fact, the staircase function in the above figure, $s(x) = \sum_i \Delta y 1_{\{x : f(x) > y_i\}}$, is a simple function, and $\int s = \sum_i \mu(\{x : f(x) > y_i\})\Delta y$, whose limit as $\Delta y \to 0$ defines $\int f$.

2.7 Convergence theorems

The most important convergence result is the monotone convergence theorem. If $f_k(x) \to f(x)$ for each x monotonically, i.e., $f_k(x) \le f_{k+1}(x)$ for each x, and each f_k is measurable, then $\int f_k \to \int f$. In order to prove this result, it is easy to prove that $\int f_k \le \int f$ because of monotonicity. Then for each simple function below f, such as the staircase function, we can also prove that eventually $\int f_k \ge \int s$ because $f_k(x) \to f(x)$ for each x whereas s is below f. Thus we have $\int f_k \to \int f$. So the simple functions play an important role in the proof

of the monotone convergence theorem, because they play a crucial role in defining the Lebesgue integral. Another two similar results are dominated convergence theorem and Fatou's lemma.

Lebesgue integral gives us a mathematical system where the integrals are well defined and well behaved. Compared to Riemann integral, it is more general and self-consistent, even though they agree with each other for all practically relevant cases.

3 Univariate distribution and expectation

3.1 Discrete random variable, expectation, long run average

In Example 1, X has a uniform distribution over $\{1,2,...,6\}$. $P(X=x)=1/6$ for $x \in \{1,2,...,6\}$. In general, $p(x)=P(X=x)$ is called the probability mass function. We use capital letters for random variables, and lower case letters for the values that the random variables can take. We can arrange $p(x)$ into a table.

Given $p(x)$, the additivity of probability tells us that

$$P(X \in A) = P(A) = \sum_{x \in A} p(x).$$

The expectation of X is defined as

$$\mathbb{E}(X) = \sum_x xp(x).$$

It can be interpreted as long run average or center of fluctuation. Let $X_i \sim p(x)$ independently for $i=1,...,n$, then according to the law of large numbers that we shall prove,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}(X).$$

The interpretation of this result is that the frequency of $X_i = x$ approaches $p(x)$, i.e., the number of times $X_i = x$ is about $np(x)$. Thus the sum is $\sum_x np(x)x$, and the average is $\sum_x xp(x)$, which is $\mathbb{E}(X)$.

More generally,

$$\mathbb{E}[h(X)] = \sum_x h(x)p(x).$$

It can be interpreted as long run average of $h(X)$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \to \mathbb{E}[h(X)].$$

We define the variance as the average squared deviation from the center

$$\mathrm{Var}(X) = \mathbb{E}[X - \mathbb{E}(X)]^2,$$

and

$$\mathrm{Var}[h(X)] = \mathbb{E}[h(X) - \mathbb{E}[h(X)]]^2.$$

It tells us the magnitude of fluctuation, variability, volatility, and spread of the distribution.

We can show that

$$\begin{aligned} \mathbb{E}(aX+b) &= \sum_x (ax+b)p(x) \\ &= a\sum_x xp(x) + b\sum_x p(x) \\ &= a\mathbb{E}(X) + b, \end{aligned}$$

$$\begin{aligned}\mathrm{Var}(aX+b) &= \mathbb{E}[(aX+b)-\mathbb{E}(aX+b)]^2 \\ &= \mathbb{E}[(aX+b)-(a\mathbb{E}(X)+b)]^2 \\ &= \mathbb{E}[a^2(X-\mathbb{E}(X))^2] = a^2\mathrm{Var}(X).\end{aligned}$$

There is a short cut formula for $\mathrm{Var}(X)$:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[X-\mathbb{E}(X)]^2 \\ &= \mathbb{E}[X^2-2X\mathbb{E}(X)+\mathbb{E}(X)^2] \\ &= \mathbb{E}(X^2)-2\mathbb{E}(X)^2+\mathbb{E}(X)^2 \\ &= \mathbb{E}(X^2)-\mathbb{E}(X)^2.\end{aligned}$$

Let $h(X)=aX+b$, then $\mathbb{E}[h(X)]=h(\mathbb{E}(X))$. Let $h(X)=X^2$, then $\mathbb{E}(h(X))\geq h(\mathbb{E}(X))$, which is the Jensen inequality for convex function h.

Let $\mu=\mathbb{E}(X)$ and $\sigma^2=\mathrm{Var}(X)$. Let $Z=(x-\mu)/\sigma$, then

$$\mathbb{E}(Z)=\mathbb{E}(X-\mu)/\sigma=(\mathbb{E}(X)-\mu)/\sigma=0,$$

$$\mathrm{Var}(Z)=\mathrm{Var}(X-\mu)/\sigma^2=\mathrm{Var}(X)/\sigma^2=1.$$

3.2 Continuous random variable, basic event, discretization

For a continuous random variable X, such as the height of a random person, the basic event cannot be $X=x$, because its probability is 0, just as the length of a point is 0. Instead, we use $X\in(x,x+\Delta x)$ as the basic event. This is like discretization of the continuous range into a discrete collection of bins $(x,x+\Delta x)$. For infinitesimal $\Delta x\to 0$, let

$$P(X\in(x,x+\Delta x))=f(x)\Delta x,$$

or

$$f(x)=\frac{P(X\in(x,x+\Delta x))}{\Delta x},$$

then we can replace $p(x)=P(X=x)$ for the discrete random variable by $P(X\in(x,x+\Delta x))=f(x)\Delta x$ for the continuous random variable, so that

$$\begin{aligned}P(X\in A)=P(A) &= \sum_x p(x) \text{ (discrete case)} \\ &= \sum_{(x,x+\Delta x)\subset A} f(x)\Delta x\to\int_A f(x)dx.\end{aligned}$$

For an interval $A=(a,b)$, $P(A)$ is the area under the curve $f(x)$ within the interval A. Similarly, we can define expectation as

$$\begin{aligned}\mathbb{E}(X) &= \sum_x xp(x) \text{ (discrete case)} \\ &= \sum_{(x,x+\Delta x)} xf(x)\Delta x\to\int xf(x)dx.\end{aligned}$$

Moreover,

$$\mathbb{E}(h(X))=\sum_{(x,x+\Delta x)} h(x)f(x)\Delta x\to\int h(x)f(x)dx.$$

The definition of $\mathrm{Var}(X)$ and $\mathrm{Var}[h(X)]$, and the properties of $\mathbb{E}(aX+b)$ and $\mathrm{Var}(aX+b)$ remain the same.

The long-run average interpretation also remains the same. Let $X_i \sim f(x)$ independently for $i = 1, ..., n$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}(X).$$

The interpretation is that the frequency that $X_i \in (x, x+\Delta x)$ is $f(x)\Delta x$, i.e., the number of X_i in $(x, x+\Delta x)$ is about $nf(x)\Delta x$. For all those X_i within $(x, x+\Delta x)$, they are approximately equal to x, and their sum is $xnf(x)\Delta x$. The total sum is $\sum xnf(x)\Delta x$, and the average is $\sum h(x)f(x)\Delta x$, which approaches $\mathbb{E}(X)$ as $\Delta x \to 0$. The same with the interpretation of $\mathbb{E}[h(X)]$ as the long run average of $h(X_i)$.

3.3 How to think about density

Consider the scatter plot of the sample $X_1, ..., X_n$ for a large n, where each X_i is a point on the real line. The points will be denser around those x where $f(x)$ is high. In fact, $f(x)$ = number of points in $(x, x+\Delta x)/n/\Delta x$.

We may also plot a histogram for $X_1, ..., X_n$, by distributing them into a number of bins. The number of X_i in the bin $(x, x+\Delta x)$ is about $nf(x)\Delta x$, i.e., the proportion of those X_i in $(x, x+\Delta x)$ is $f(x)\Delta x$.

We may also consider a population Ω. We can again consider the scatter plot of the population $\{X(\omega), \omega \in \Omega\}$ or the histogram of this population. Consider each $X(\omega)$ as a point on the real line, $f(x)$ describes the population density of the points at x. $f(x)$ is also the histogram of the population, where the proportion in the bin $(x, x+\Delta x)$ is $f(x)\Delta x$.

In general, we can think of either a population $\{X(\omega), \omega \in \Omega\}$ or a large sample $\{X_i, i = 1, ..., n\}$. Both can make the probability distribution more tangible.

With the above interpretation, $\mathbb{E}(X)$ can be considered the population average of $X(\omega)$ for all $\omega \in \Omega$ or the long run average of $\{X_i, i = 1, ..., n\}$.

We can consider a probability density as a population of points. When we sample from this probability density, it is as if we randomly sample a point from this population.

$\mathbb{E}(X)$ can be considered the mass center of the large sample of points or the population of points.

We may also consider a large sample or a population of points uniformly distributed within the region under the curve $f(x)$. Then the distribution of the horizontal coordinates of these points is $f(x)$, or in other words, if we project these points onto the x-axis, we get a large sample or a population of points on the x-axis whose density is $f(x)$.

In quantum physics, $f(x) = |\phi(x)|^2$ for an electron is the probability density function of the position of the electron when we make an observation. Physicists intuitively think of $f(x)$ as a cloud of points whose density is $f(x)$.

In statistical physics, the state of a system that consists of a large number of elements is assumed to follow a distribution. Physicists call such a distribution an ensemble, which has the same meaning as population.

In physics, the expectation $\mathbb{E}(h(X))$ is denoted as $\langle h(X) \rangle$.

3.4 Existence of probability density function

For a random variable X, let $\sigma(X)$ be the σ-algebra generated by X, i.e., the collection of all the meaningful statements about X, such as $X \in (a, b)$. Let $P(A) = P(X \in A)$ for $A \in \sigma(X)$. $P(A)$ is a measure. Let μ be the usual Lebesgue measure, i.e., length. If P is absolutely continuous with respect to μ, i.e., if $\mu(A) = 0$ then $P(A) = 0$. Then according to the Radon-Nikodym theorem, there exists a density function, such as $P(A) = \int_A f d\mu$. We can write $f = dP/d\mu$.

3.5 Cumulative density

The cumulative density function $F(x) = P(X \leq x)$. It is defined for both discrete and continuous random variables, but it is more natural for continuous random variables.

If x is the GRE score, then $F(x)$ is the percentile of the score x. For $u \in [0,1]$, $x = F^{-1}(u)$ is the quantile, i.e., the original score whose percentile is u, i.e., $F(x) = u$.

$F(x)$ is a monotone non-decreasing function in x.

$f(x) = F'(x)$, because

$$\begin{aligned} F'(x) &= \frac{F(x+\Delta x) - F(x)}{\Delta x} \\ &= \frac{P(X \leq x+\Delta x) - P(X \leq x)}{\Delta x} \\ &= \frac{P(X \in (x, x+\Delta x))}{\Delta x} = f(x). \end{aligned}$$

3.6 Uniform distribution

Let $U \sim \text{Uniform}[0,1]$, i.e., the density of U is $f(u) = 1$ for $u \in [0,1]$, and $f(u) = 0$ otherwise.

(1) Calculate $F(u) = P(U \leq u)$.

$$F(u) = P(U \leq u) = \begin{cases} 0 & 0 < u \\ u & 0 \leq u \leq 1 \\ 1 & u > 1 \end{cases}$$

(2) Calculate $\mathbb{E}(U)$, $\mathbb{E}(U^2)$, and $\text{Var}(U)$.

$$\mathbb{E}(U) = \int_0^1 u du = \frac{1}{2}$$

$$\mathbb{E}(U^2) = \int u^2 du = \frac{1}{3}$$

$$\text{Var}(U) = \mathbb{E}(U^2) - (\mathbb{E}(U))^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

3.7 Inversion method

The inversion method is for generating a random variable $X \sim F(x)$. We only need to generate $U \sim \text{Uniform}[0,1]$, and let $X = F^{-1}(U)$. This is because

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(F^{-1}(U) \leq x) \\ &= P(U \leq F(x)) = F(x). \end{aligned}$$

Intuitively, we can imagine mapping the equally spaced points in $[0,1]$ on the vertical axis to the points on the horizontal axis via $u = F(x)$ or $x = F^{-1}(u)$. The density is determined by the slope of $F(x)$, which is the density $f(x)$ of X.

We may also consider the uniform distribution as a population of points, and $F(x)$ describes anther population of points. We can map the points in the uniform distribution to the points in the F distribution by preserving the order. For a point u, the proportion of points before it is u. For a point x, the proportion of points before it is $F(x)$. In order to preserve the order when we map u to x, we want $u = F(x)$, so $x = F^{-1}(u)$.

We may also consider $U \sim \text{uniform}[0,1]$ as the rank or percentile of a random person. $F^{-1}(U)$ is the original score of a random person.

As an example, if we want to generate $X \sim \text{Exponential}(1)$, where $F(x) = 1 - e^{-x}$. Then $F^{-1}(u)$ can be obtained by solving $1 - e^{-x} = u$, so $x = -\log(1-u)$. So we can generate $X = -\log(1-U)$ or simply $X = -\log U$.

3.8 Transformation

Suppose $X \sim f_X(x)$, and let $Y = h(X)$. In order to derive the density of Y, i.e., $f_Y(y)$, we can use two methods. One is via the cumulative density.

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(h(X) \le y) \\ &= p(X \in \{x : h(x) \le y\}). \end{aligned}$$

Then we calculate $f_Y(y) = F_Y'(y)$.

The other method is to calculate the density directly,

$$\begin{aligned} f_Y(y) &= \frac{P(Y \in (y, y+\Delta y))}{\Delta y} \\ &= \frac{P(h(X) \in (y, y+\Delta y))}{\Delta y} \\ &= \frac{P(X \in \{x : h(x) \in (y, y+\Delta y))}{\Delta y}. \end{aligned}$$

To be more concrete, consider $Y = h(X) = aX + b$, with $a > 0$. Then the event

$$\{x : h(x) \le y\} = \{x : x \le (y-b)/a\}.$$

Thus $F_Y(y) = F_X((y-b)/a)$, so $f_Y(y) = f_X((y-b)/a)/a$. For general $a \neq 0$, it should be $f_Y(y) = f_X((y-b)/a)/|a|$.

Or with $a > 0$,

$$\{x : h(x) \in (y, y+\Delta y)\} = \{x : x \in ((y-b)/a, (y-b)/a + \Delta y/a)\}.$$

So

$$f_Y(y) = f_X((y-b)/a)\Delta y/a/\Delta y = f_X((y-b)/a)/a.$$

Again in general, $f_Y(y) = f_X((y-b)/a)/|a|$.

For a monotone increasing $h(x)$, let $x = g(y)$ be the inverse of h, then

$$\{x : h(x) \le y\} = \{x : x \le g(y)\}.$$

So $F_Y(y) = F_X(g(y))$, and $f_Y(y) = f_X(g(y))g'(y)$. Or

$$\{x : h(x) \in (y, y+\Delta y)\} = \{x : x \in (g(y), g(y+\Delta y))\},$$

so

$$f_Y(y) = \frac{f_X(g(y))(g(y+\Delta y) - g(y))}{\Delta y} = f_X(g(y))g'(y).$$

In general, for a monotone $h(x)$, $f_Y(y) = f_X(g(y))|g'(y)|$.

The intuition is as follows. We can think about a distribution $f_X(x)$ as a population (or a large sample) of points on the x-axis. If we map these points to a population of points on the y-axis via the transformation $y = h(x)$, then the density will change. If the slope of $h(x)$ is small at x, we will increase the density. If the slope is big, we will decrease the density. This is because $h(x)$ maps the neighborhood $(x, x+\Delta x)$ to $(y, y+\Delta y)$, where $y = h(x)$, and $\Delta y = h'(x)\Delta x$. But the number of points in $(x, x+\Delta x)$ is the same as the number of points in $(y, y+\Delta y)$ in the mapping. So the density changes because of the change of the size of neighborhood from Δx to Δy. This is demonstrated in the following plot:

Another intuition is to think about $f_X(x)$ as a histogram of small balls with a large number of bins $(x, x+\Delta x)$, where each small ball is a member of the population or the big sample. Under the transformation $y = h(x)$, we change the bin $(x, x+\Delta x)$ to a bin $(y, y+\Delta y)$, where $\Delta y/\Delta x = h'(x)$. If $\Delta y > \Delta x$, i.e., we stretch the bin $(x, x+\Delta x)$, then the histogram in this bin will drop. If $\Delta y < \Delta x$, i.e., we squeeze the bin $(x, x+\Delta x)$, then the histogram in this bin will raise. The factor of change is $\Delta x/\Delta y$.

A symbolic way of memorizing the above formula is

$$\begin{aligned} X &\sim f_X(x)dx \\ &\sim f_X(g(y))dg(y) \\ &\sim f_X(g(y))|g'(y)|dy \\ &\sim f_Y(y)dy \sim Y. \end{aligned}$$

This formula can go both directions. If we notice $y \sim f(g(y))dg(y)$, we can then claim that $X = g(Y) \sim f(x)dx$.

For multivariate $y = h(x)$, or $x = g(y)$, where both x and y are d-dimensional vectors, then $|g'(y)|$ is the absolute value of the determinant of $g'(y)$, which is a $d \times d$ matrix whose (i, j) entry is $\partial x_i/\partial y_j$. Let D_y be a local region around y. Suppose the points in D_y are mapped to the points in D_x, which is the local region around x, by $x = g(y)$. Then

$$\begin{aligned} f_Y(y) &= \frac{P(Y \in D_y)}{|D_y|} \\ &= \frac{P(X \in D_x)}{|D_y|} \\ &= \frac{f_X(x)|D_x|}{|D_y|} = f_X(x)|g'(y)|. \end{aligned}$$

That is $|g'(y)|$ is the ratio between the volume of D_x and the volume of D_y.

For a linear transformation $x = Ay$, it maps the unit cube in y to the parallelogram formed by the column vectors of A. So the determinant $|A|$ is the volume of this parallelogram.

3.9 Polar method for normal random variable

Consider generating $(X, Y) \sim \mathrm{N}(0, 1)$ independently. Then

$$f(x, y) = \frac{1}{2\pi} \exp(-(x^2 + y^2)/2).$$

If we map (x, y) to (r, θ) via the polar transformation $x = r\cos\theta$ and $y = r\sin\theta$, then

$$f(r, \theta) = \frac{1}{2\pi} e^{-r^2/2} r,$$

where $|D_{x,y}|/|D_{r,\theta}| = r$. Thus R and Θ are independent, with $\Theta \sim \mathrm{Uniform}[0, 2\pi]$ and $R \sim e^{-r^2/2} dr^2/2$, so $T = R^2/2 \sim e^{-t}dt$. Thus we can generate $T = -\log U_1$, and $\Theta = 2\pi U_2$, and calculate X and Y.

Another way to think about it is to consider the ring $\{(x,y) : \sqrt{x^2+y^2} \in (r, r+\Delta r)$. The area of this ring is $2\pi r\Delta r$. Then

$$P(R \in (r, r+\Delta r)) = \frac{1}{2\pi} e^{-r^2/2} 2\pi r = e^{-r^2/2} r,$$

and R is independent of Θ because the distribution is isotropic.

This method of generating normal random variables is called the polar method. This is the way we prove that the normal density integrates to 1.

3.10 Counting techniques

(1) *Number of labeled pairs*. If Experiment 1 has n_1 outcomes, and Experiment 2 has n_2 outcomes, then the number of outcomes from the two experiments is $n_1 \times n_2$. We can generalize this result to the number of labeled k tuples.

As an example of labeled pairs, if we roll a die twice, let X_1 be the number of the first throw, and X_2 the number of the second throw. Then the number of all possible (X_1, X_2) is 6×6. Here X_1 and X_2 are labeled by the subscripts 1 and 2. $(X_1 = 1, X_2 = 2)$ is different from $(X_1 = 2, X_2 = 1)$.

You can understand this rule by a two way table or by a branching tree diagram.

(2) *Number of permutations*. If there are n cards, and we sequentially pick $k \leq n$ cards without replacement (i.e., do not put back), where order matters. Then the number of possible sequences is $P_{n,k} = n(n-1)...(n-k+1)$. A special case is $k = n$, which is the number of ways to permute n different objects, which is $n!$.

(3) *Number of combinations*. If there are n different balls, and we pick $k \leq n$ balls without replacement, where order does not matter. Then the number of possible combinations is $\binom{n}{k} = P_{n,k}/k! = n!/(k!(n-k)!)$. The reason is that each combination corresponds to $k!$ permutations.

Coin flipping. If we flip a fair coin n times independently. Let X be the number of heads. Then $P(X = k) = \binom{n}{k}/2^n$. To be more precise, let Ω be a sequence, and let $X(\omega)$ be the number of heads in sequence Ω. Then $A = \{\omega : X(\omega) = k\}$ is the set of sequences with exactly k heads. $|A| = \binom{n}{k}$, because in order to generate a sequence with exactly k heads, we may list n blanks, and choose k blanks to fill in heads, and fill the rest $n-k$ blanks by tails. $|\Omega| = 2^n$.

Hyper-geometric distribution. Suppose there is an urn with R red balls, and B black balls. Let $N = R+B$. Suppose we randomly pick n balls without replacement. Let X be the number of red balls that we pick. Then $P(X = r) = \binom{R}{r}\binom{B}{b}/\binom{N}{n}$, where $r \leq R$, $r \leq n$, and $b = n-r$, and $b \leq B$. Let $p = R/N$. Suppose we fix p and let R and N go to infinity. Then $P(X = r) \to \binom{n}{r} p^r (1-p)^{n-r}$, which is the binomial probability as we shall study later.

3.11 Bernoulli

For $Z \sim \text{Bernoulli}(p)$, $Z \in \{0,1\}$, $P(Z = 1) = p$ and $P(Z = 0) = 1-p$. Then

$$\mathbb{E}(Z) = 0 \times (1-p) + 1 \times p = p.$$

$$\begin{aligned} \text{Var}(X) &= (0-p)^2 \times (1-p) + (1-p)^2 \times p \\ &= p(1-p)[p + (1-p)] = p(1-p). \end{aligned}$$

3.12 Binomial

For $X \sim \text{Binomial}(n,p)$, X is the number of heads if we flip the coin n times independently, where the probability of head in each flip is p.

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

$\binom{n}{k}$ is the number of sequences with exactly k heads. The reason is that, to produce such a sequence, we need to choose k positions from the n positions to place heads, and we place tails in the remaining $n-k$ positions. $p^k(1-p)^{n-k}$ is the probability of each sequence with exactly k heads. The reason is that there are k heads and $n-k$ tails, and the flips are independent. The probability of observing each head is p and the probability of observing each tail is $1-p$.

A useful representation of X is

$$X = Z_1 + Z_2 + ... + Z_n,$$

where $Z_i \sim \text{Bernoulli}(p)$ independently. The reason is that the sum of binary variables equals to the number of 1's among these binary variables. Thus

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(Z_i) = np.$$

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(Z_i) = np(1-p).$$

X/n is the frequency of heads. $\mathbb{E}(X/n) = \mathbb{E}(X)/n = p$. $\text{Var}(X/n) = \text{Var}(X)/n^2 = p(1-p)/n$. $\text{Var}(X) \to 0$ as $n \to \infty$. So $X/n \to p$. That is, long run frequency converges to probability.

The name "Binomial" comes from the binomial formula:

$$(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}.$$

The above formula can be understood in the same way as we understand the distribution. $\prod_{i=1}^{n}(p_i+q_i)$ can be expanded into 2^n sequences. The sequence with exactly k p's is $\binom{n}{k}$. Assuming $p_i = p$ and $q_i = q$ for all i, then each such sequence is $p^k q^{n-k}$.

The Binomial distribution can also be understood in terms of quincunx or Galton's board, which his related to the Pascal triangle that organize the numbers of n,k.

3.13 Normal approximation

Suppose $X \sim \text{Binomial}(n,1/2)$, i.e., the number of heads from n independent flips of a fair coin. Then $\mu = \mathbb{E}(X) = n/2$, $\sigma^2 = \text{Var}(X) = n/4$, $\sigma = SD(X) = \sqrt{n}/2$. Let $Z = (X-\mu)/\sigma$, then $\mathbb{E}(Z) = 0$, $\text{Var}(Z) = 1$, no matter what n is.

We know that

$$P(X=k) = \frac{\binom{n}{k}}{2^n} = \frac{n!}{k!(n-k)!2^n},$$

but for large n, it is hard to use this formula to calculate the probabilities. We can find asymptotic approximations using the Stirling formula

$$n! \sim \sqrt{2\pi n} n^n e^{-n},$$

where $a(n) \sim b(n)$ means $\lim_{n\to\infty} a(n)/b(n) = 1$. Our plan is to first calculate $P(X = n/2)$. Then we calculate $P(X = n/2+d)/P(X = n/2)$. The scale of d is chosen to be σ, i.e., $d = z\sigma = z\sqrt{n}/2$ where z is

fixed, and $n \to \infty$. The reason we consider such deviations is that the scale of $Z = (X-\mu)/\sigma$ is independent of n, i.e., X can be represented by $X = \mu + Z\sigma$ for a Z whose expectation is 0 and whose variance is 1 no matter how large n is.

$$\begin{aligned} P(X = n/2) &\sim \frac{n!}{(n/2)!^2 2^n} \\ &\sim \frac{\sqrt{2\pi n} n^n e^{-n}}{(\sqrt{2\pi(n/2)}(n/2)^{n/2})^2 2^n} \\ &\sim \frac{1}{\sqrt{2\pi}} \frac{2}{\sqrt{n}}. \end{aligned}$$

Let $k = n/2 + z\sqrt{n}/2 = n/2 + d$.

$$\begin{aligned} \frac{P(X = n/2+d)}{P(X = n/2)} &= \frac{(n/2)!(n/2)!}{(n/2+d)!(n/2-d)!} \\ &= \frac{(n/2)(n/2-1)...(n/2-(d-1))}{(n/2+1)(n/2+2)...(n/2+d)} \\ &= \frac{1(1-2/n)(1-2\times 2/n)...(1-(d-1)\times 2/n)}{(1+2/n)(1+2\times 2/n)...(1+d\times 2/n)} \\ &= \frac{(1-\delta)(1-2\delta)...(1-(d-1)\delta)}{(1+\delta)(1+2\delta)...(1+d\delta)} \\ &\to \frac{e^{-\delta}e^{-2\delta}...e^{-(d-1)\delta}}{e^{\delta}e^{2\delta}...e^{d\delta}} \\ &= \frac{e^{-(1+2+...+(d-1))\delta}}{e^{(1+2+...+d)\delta}} \\ &= \frac{e^{-d(d-1)\delta/2}}{e^{d(d+1)\delta/2}} \\ &= e^{-[d(d-1)/2+d(d+1)/2]\delta} = e^{-d^2\delta} \\ &= e^{-(z\sqrt{n}/2)^2(2/n)} = e^{-\frac{z^2}{2}}, \end{aligned}$$

where $\delta = 2/n$, and $d = z\sqrt{n}/2$. Thus

$$\begin{aligned} P(X = n/2 + z\sqrt{n}/2) &= P(X = \mu + z\sigma) \\ &\sim \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \frac{2}{\sqrt{n}} = f(z)\Delta z, \end{aligned}$$

where $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ and $\Delta z = \frac{2}{\sqrt{n}}$. Thus with $\mu = n/2$, $\sigma = \sqrt{n}/2$, and $Z = (X-\mu)/\sigma$, we have

$$\begin{aligned} P(X \in [\mu + a\sigma, \mu + b\sigma]) &= P(Z \in [a,b]) \\ &= \sum_{k=\mu+a\sigma}^{\mu+b\sigma} P(X = k) \\ &= \sum_{z\in[a,b]} P(X = \mu + z\sigma) \\ &= \sum_{z\in[a,b]} f(z)\Delta z \to \int_a^b f(z)dz = \Phi(b) - \Phi(a), \end{aligned}$$

where the space between two consecutive values of $z = (k-\mu)/\sigma$ is $1/\sigma = 2/\sqrt{n} = \Delta z$, and $\Phi(z) = \int_{-\infty}^{z} f(z)dz$ is the cumulative density function of Z. So if $X \sim \text{Binomial}(n, 1/2)$, and $Z = (X-\mu)/\sigma$, then in the limit, $Z \sim \text{N}(0,1)$, i.e., the standard normal distribution, whose density function is $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$.

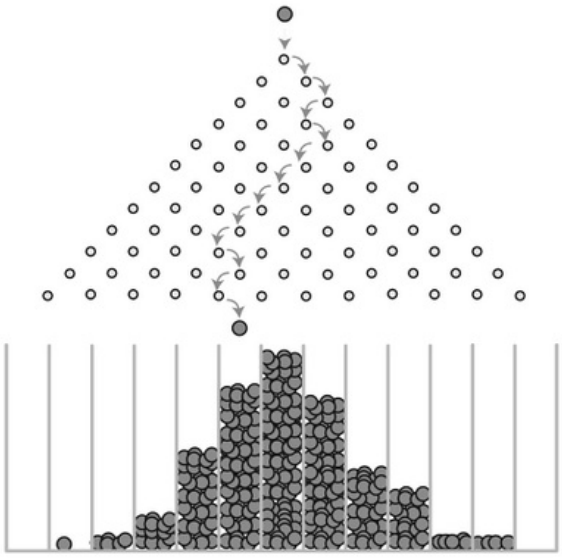

Figure 3: Galton's board. Source: web.

The above normal approximation can be illustrated by the quincunx or Galton's board.

For general $p \in (0,1)$, $\mu = np$ and $\sigma^2 = npq$, $q = 1-p$. Let $k = \mu + z\sigma = np + z\sqrt{npq} = np + d$. Neglecting $o(1/n)$ terms, we have

$$\begin{aligned} P(X = np+d) &= \frac{n!}{(np+d)!(nq-d)!} p^{np+d} q^{nq-d} \\ &= \frac{\sqrt{2\pi n} n^n e^{-n} p^{np+d} q^{nq-d}}{\sqrt{2\pi(np+d)}(np+d)^{np+d} e^{-np-d} \sqrt{2\pi(nq-d)}(nq-d)^{nq-d} e^{-nq+d}} \\ &= \frac{\sqrt{n}}{\sqrt{2\pi(np+d)(nq-d)}} (\frac{np}{np+d})^{np+d} (\frac{nq}{nq-d})^{nq-d}. \end{aligned}$$

Take log, and use Taylor expansion $\log(1+\delta) = \delta - \delta^2/2 + O(\delta^3)$, we have

$$\begin{aligned} \log\left[(\frac{np}{np+d})^{np+d}(\frac{nq}{nq-d})^{nq-d}\right] &= -(np+d)\log\left(1+\frac{d}{np}\right) - (nq-d)\log\left(1-\frac{d}{nq}\right) \\ &= -(np+d)\left[\frac{d}{np} - \frac{1}{2}\left(\frac{d}{np}\right)^2\right] - (nq-d)\left[-\frac{d}{nq} - \frac{1}{2}\left(\frac{d}{nq}\right)^2\right] \\ &= -\frac{d^2}{2}\left[\frac{1}{np} + \frac{1}{nq}\right] = -\frac{z^2}{2} \end{aligned}$$

Moreover,

$$\frac{\sqrt{n}}{\sqrt{2\pi(np+d)(nq-d)}} \sim \frac{1}{\sqrt{2\pi npq}}.$$

Thus

$$P(X = u + \sigma z) \sim \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \Delta z = f(z)\Delta z,$$

where $\Delta z = \Delta x/\sigma = 1/\sqrt{npq}$. Thus $P(Z \in (a,b)) = \sum_{z\in(a,b)} f(z)\Delta z \to \int_a^b f(z)dz$.

Let $Z \sim \mathrm{N}(0,1)$, i.e., the density of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

(1) Calculate $\mathbb{E}(Z)$, $\mathbb{E}(|Z|)$, $\mathbb{E}(Z^2)$, and $\mathrm{Var}(Z)$.

$$\begin{aligned}
\mathbb{E}(|Z|) &= 2\int_0^\infty \frac{1}{\sqrt{2\pi}} z e^{-\frac{z^2}{2}} dz \\
&= -2\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}|_0^\infty \\
&= -2\frac{1}{\sqrt{2\pi}}(-1) = \frac{2}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}
\end{aligned}$$

$\mathbb{E}(Z) = 0$ because the density is symmetric around 0.

$$\begin{aligned}
\mathbb{E}(Z^2) &= \int_{-\infty}^\infty z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty (-z) de^{-\frac{z^2}{2}} \\
&= \frac{1}{\sqrt{2\pi}} (-z e^{-\frac{z^2}{2}}|_{-\infty}^\infty - \int_{-\infty}^\infty e^{-\frac{z^2}{2}} d(-z)) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{z^2}{2}} dz = 1
\end{aligned}$$

$$\mathrm{Var}(Z) = \mathbb{E}(Z^2) - (\mathbb{E}(Z))^2 = 1$$

(2) Let $X = \mu + \sigma Z$. Find the density of X. Calculate $\mathbb{E}(X)$ and $\mathrm{Var}(X)$.

$$Z = \frac{X - \mu}{\sigma}.$$

Using the symbolic formula,

$$\begin{aligned}
Z &\sim f_Z(z)dz = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) d\left(\frac{x-\mu}{\sigma}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= f_X(x)dx \sim X.
\end{aligned}$$

Moreover,

$$\mathbb{E}(X) = E(\mu + \sigma Z) = \mu + \sigma\mathbb{E}(Z) = \mu.$$
$$\mathrm{Var}(X) = \mathrm{Var}(\mu + \sigma Z) = \sigma^2 \mathrm{Var}(Z) = \sigma^2.$$

3.14 Geometric

For $T \sim \text{Geometric}(p)$, T is the number of flips to get the first head, if we flip a coin independently and the probability of getting a head in each flip is p. $P(T=k) = (1-p)^{k-1}p$, where $k = 1, 2, ...$ The reason is that the first $k-1$ flips are tails, and the k-th one is head. Let $q = 1-p$,

$$\begin{aligned}
\mathbb{E}(T) &= \sum_{k=1}^{\infty} kP(T=k) \\
&= \sum_{k=1}^{\infty} kq^{k-1}p = p\sum_{k=1}^{\infty} \frac{d}{dq}q^k \\
&= p\frac{d}{dq}\sum_{k=1}^{\infty} q^k = p\frac{d}{dq}\left(\frac{1}{1-q} - 1\right) \\
&= p\frac{1}{(1-q)^2} = \frac{1}{p}.
\end{aligned}$$

This result is easy to understand. Suppose $p = 1/3$. This means that, on average, 1 out of 3 times we get a head. So on average, we need to flip 3 times to get a head, and $3 = 1/(1/3) = 1/p$. If $p = 1/10$, then on average, 1 out of 10 times we get a head, so on average we need to flip 10 times to get a head.

3.15 Poisson process

Things happen over continuous time, so we often need to study continuous time processes. The way to study such processes is very much like making a movie. When we see a movie, we get an impression (or illusion) that things happen continuously. But actually the movie theater shows a discrete sequence of frames, for instance, 24 frames a second. Similarly, we can divide the time domain into small periods of duration Δt, so we get $[0, \Delta t]$, $[\Delta t, 2\Delta t]$, $[2\Delta t, 3\Delta t]$, ...Then we can model what happens with each period. Then we can calculate the long term consequences.

For instance, we may assume that within each period we flip a coin independently. Let $p = h(\Delta t)$ be the probability of getting a head. By Taylor expansion,

$$h(\Delta t) = h(0) + h'(0)\Delta t + h''(0)\Delta t^2/2 + ...$$

Clearly as $\Delta t \to 0$, $p \to 0$. So we want $h(0) = 0$. Also, for infinitesimal Δt, we may discard the high order terms, so we may simply assume that $p = \lambda \Delta t$, where λ can be interpreted as rate or intensity of getting a head (which, in real life, may be the occurrence of an earthquake or the decay of a particle). For instance, λ = once every 10 years, or 3 times per second.

Let T be the waiting time until the first head.

$$P(T \in (t, t+\Delta t)) = (1-\lambda\Delta t)^{t/\Delta t}\lambda\Delta t,$$

because there are $t/\Delta t$ flips before t, and they are all tails. The flip in $[t, t+\Delta t]$ results in a head. To simplify the expression, we use the result $e^x = 1 + x + x^2/2 + ...$ from Taylor expansion. If x is infinitesimally small, then we can approximately write $1 + x \to e^x$. Thus

$$\frac{P(T \in (t, t+\Delta t))}{\Delta t} \to (e^{-\lambda\Delta t})^{t/\Delta t}\lambda = \lambda e^{-\lambda t}.$$

So in the limit, T can be considered a continuous random random, $T \sim f(t) = \lambda e^{-\lambda t}$ for $t \geq 0$. We call this distribution Exponential(λ). The survival probability

$$P(T > t) = (1-\lambda\Delta t)^{t/\Delta t} \to (e^{-\lambda\Delta t})^{t/\Delta t} = e^{-\lambda t},$$

because $T > t$ means that all the $t/\Delta t$ flips before time t are tails.

We can write $T = \tilde{T}\Delta t$, where $\tilde{T} \sim \text{Geometric}(p = \lambda\Delta t)$. Then

$$\mathbb{E}(T) = \mathbb{E}(\tilde{T})\Delta t = \frac{1}{p}\Delta t = \frac{1}{\lambda\Delta t}\Delta t = 1/\lambda.$$

Let $T \sim \text{Exponential}(\lambda)$, i.e., the density of T is $f(t) = \lambda e^{-\lambda t}$ for $t \geq 0$, and $f(t) = 0$ for $t < 0$.

(1) Calculate $F(t) = P(T \leq t)$. Find t so that $F(t) = 1/2$.

$$\begin{aligned} F(t) &= \int_0^t f(t)dt = \int_0^t \lambda e^{-\lambda t}dt \\ &= -e^{-\lambda t}|_0^t = 1 - e^{-\lambda t} \end{aligned}$$

$$F(t) = \frac{1}{2} = 1 - e^{-\lambda t}$$

$$t = \frac{\log 1/2}{-\lambda} = \frac{\log 2}{\lambda}$$

(2) Calculate $\mathbb{E}(T)$, $\mathbb{E}(T^2)$, and $\text{Var}(T)$.

$$\begin{aligned} \mathbb{E}(T) &= \int_0^\infty t\lambda e^{-\lambda t}dt \\ &= -\int_0^\infty t de^{-\lambda t} \\ &= -(te^{-\lambda t}|_0^\infty - \int_0^\infty e^{-\lambda t}dt) \\ &= -(0 - 0 + \frac{1}{\lambda}e^{-\lambda t}|_0^\infty) = \frac{1}{\lambda} \end{aligned}$$

$$\begin{aligned} \mathbb{E}(T^2) &= \int_0^\infty t^2\lambda e^{-\lambda t}dt \\ &= -\int_0^\infty t^2 de^{-\lambda t} = -(t^2e^{-\lambda t}|_0^\infty - \int_0^\infty e^{-\lambda t}2tdt) \\ &= 2\int_0^\infty e^{-\lambda t}tdt = \frac{2}{\lambda}\int_0^\infty t\lambda e^{-\lambda t}dt \\ &= \frac{2}{\lambda}\frac{1}{\lambda} = \frac{2}{\lambda^2} \end{aligned}$$

$$\text{Var}(T) = \mathbb{E}(T^2) - (\mathbb{E}(T))^2 = \frac{1}{\lambda^2}$$

The Poisson process is a model for the occurrences of rare events, because the probability of occurrence in each small time period is $\lambda\Delta t$. It is a very important model in stochastic processes (or random processes).

Let X be the number of heads within $[0,t]$, then $X \sim \text{Binomial}(n = t/\Delta t, p = \lambda\Delta t)$. So

$$\begin{aligned} P(X = k) &= \frac{n(n-1)...(n-k+1)}{k!}p^k(1-p)^{n-k} \\ &= \frac{t/\Delta t(t/\Delta t - 1)...(t/\Delta t - k + 1)}{k!}(\lambda\Delta t)^k(1-\lambda\Delta t)^{t/\Delta t - k} \\ &= \frac{t(t-\Delta t)(t-2\Delta t)...(t-(k-1)\Delta t)}{k!}\lambda^k(1-\lambda\Delta t)^{t/\Delta t}(1-\lambda\Delta)^{-k} \\ &\to \frac{t^k}{k!}\lambda^k(e^{-\lambda\Delta})^{t/\Delta t} = \frac{(\lambda t)^k}{k!}e^{-\lambda t}. \end{aligned}$$

As to the expectation, since $X \sim \text{Binomial}(n = t/\Delta t, p = \lambda\Delta t)$, $\mathbb{E}(X) = np = (t/\Delta t)(\lambda\Delta t) = \lambda t$, and $\text{Var}(X) = np(1-p) \to \lambda t$. So $\lambda = \mathbb{E}(X)/t$, which is the expected number of occurrences per unit time. So λ is a rate or intensity and is in the unit of per unit time. $\mathbb{E}(T) = 1/\lambda$ is in the unit of time.

We can compute $\mathbb{E}(X)$ and $\text{Var}(X)$ directly from the probability mass function:

$$P(X = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=1}^{\infty} k\frac{(\lambda t)^k}{k!} e^{-\lambda t} \\
&= \sum_{k=1}^{\infty} \frac{(\lambda t)^k}{(k-1)!} e^{-\lambda t} \\
&= \lambda t \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \lambda t.
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1)P(X = k) = \sum_{k=2}^{\infty} k(k-1)\frac{(\lambda t)^k}{k!} e^{-\lambda t} \\
&= \sum_{k=2}^{\infty} \frac{(\lambda t)^k}{(k-2)!} e^{-\lambda t} \\
&= (\lambda t)^2 \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} = (\lambda t)^2.
\end{aligned}$$

$$\mathbb{E}(X^2) = \mathbb{E}[X(X-1)] + \mathbb{E}(X) = (\lambda t)^2 + \lambda t.$$
$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \lambda t.$$

3.16 Survival analysis

In survival analysis, λ is a function of time t, $\lambda(t)$, which is in general an increasing function. We are interested in learning whether a treatment may reduce $\lambda(t)$. We can divide the range $(0,t)$ into n equally spaced periods, with each period having duration $\Delta t = t/n$. The survival probability

$$\begin{aligned}
P(T > t) &= \prod_{i=1}^{n} (1 - \lambda(t_i)\Delta t) \\
&= \prod_{i=1}^{n} e^{-\lambda(t_i)\Delta t} \\
&= \exp\left[-\sum_{i=1}^{n} \lambda(t_i)\Delta t\right] \\
&\to \exp\left(\int_0^t \lambda(t)dt\right),
\end{aligned}$$

where t_i is within the i-th period. The density of T is

$$f(t) = \lambda(t)\exp\left(\int_0^t \lambda(t)dt\right).$$

4 Joint distribution and covariance

4.1 Joint distribution

Usually we would like to study two or more random variables together. For instance, let ω be a person from a population Ω. We may study $X(\omega)$ = eye color and $Y(s)$ = hair color together, or we study $X(\omega)$ = height and $Y(\omega)$ = weight together. For instance, we may be interested in whether eye color and hair color are somehow related to each other, or whether height and weight are somehow correlated or whether we can predict weight from height.

If both X and Y are discrete, the joint basic event is $\{X = x \text{ and } Y = y\}$, or simply written as $\{X = x, Y = y\}$. We call the probability $P(X = x, Y = y) = p(x,y)$ the joint probability mass function or joint distribution. We write $(X,Y) \sim p(x,y)$.

By the rule of total probability, i.e., additivity, the probability of marginal event $P(X = x) = \sum_y P(X = x, Y = y)$. We denote $P(X = x)$ by $p_X(x)$, where p_X is a single notation (like a name with multiple letters). Thus $p_X(x) = \sum_y p(x,y)$. We call this operation the marginalization, and $p_X(x)$ is the marginal probability mass function or marginal distribution, as $p_X(x)$ is the marginal sum of $p(x,y)$. Similarly, $p_Y(y) = \sum_x p(x,y)$.

Just like we have done before, we can interpret the above probabilities and conditional probabilities by population proportions and long run frequencies.

For population interpretation, $p(x,y)$ is the proportion of the sub-population $\{\omega : X(\omega) = x, Y(\omega) = y\}$. $p_X(x)$ is the proportion of the sub-population $\{\omega : X(\omega) = x\}$, which includes all the people with $X(\omega) = x$ and all possible values of $Y(\omega)$.

As to long run frequency, $p(x,y)$ is how often $X = x$ and $Y = y$. $p_X(x)$ is how often $X = x$, regardless of what Y is.

If both X and Y are continuous, the joint basic event is $\{X \in (x, x+\Delta x), Y \in (y, y+\Delta y)\}$. The joint density is

$$f(x,y) = \lim_{\Delta x \to 0, \Delta y \to 0} \frac{P(X \in (x, x+\Delta x), Y \in (y, y+\Delta y))}{\Delta x \Delta y}.$$

We write $(X,Y) \sim f(x,y)$.

From the joint density, we can calculate the marginal densities and conditional densities, in a similar manner as for the discrete random variables. The only difference is that we need to replace sum by integral. So $f_X(x) = \int f(x,y)dy$, $f_Y(y) = \int f(x,y)dx$. Specifically,

$$\begin{aligned} f_X(x) &= \frac{P(X \in (x, x+\Delta x))}{\Delta x} \\ &= \frac{\sum_y P(X \in (x, x+\Delta x), Y \in (y, y+\Delta y))}{\Delta x} \\ &= \frac{\sum_y f(x,y)\Delta x \Delta y}{\Delta x} \\ &= \sum_y f(x,y)\Delta y \to \int f(x,y)dy. \end{aligned}$$

We can also interpret these densities by population proportions or long run frequencies.

For the population Ω, we can think of $(X(\omega), Y(\omega))$ as the coordinate of a point. So if there are M people in Ω, then we have M points. If M is large, then these M points are like a cloud. $f(x,y)$ describes the distribution of this cloud of points or this population of points. It is very much like the distribution of a real population. For instance, the population density in NYC or LA is larger than the population density in Texas or Alaska. $f(x,y)\Delta x \Delta y$ tells us the population proportion of people or points in the cell $(x, x+\Delta x) \times (y, y+\Delta y)$.

The above interpretation can also be applied to a large sample $(X_i, Y_i) \sim f(x,y)$ independently for $i = 1, ..., n$. The scatterplot of $\{(X_i, Y_i), 1, ..., n\}$ form a cloud, whose density is described by $f(x,y)$. $f(x,y)\Delta x \Delta y$ tells us how often (frequency or proportion) $X_i \in (x, x+\Delta x)$ and $Y_i \in (y, y+\Delta y)$.

4.2 Expectation, variance, covariance

If $(X,Y) \sim p(x,y)$, then $\mathbb{E}(h(X,Y)) = \sum_x \sum_y h(x,y)p(x,y)$. If $(X,Y) \sim f(x,y)$, then $\mathbb{E}(h(X,Y)) = \int\int h(x,y)f(x,y)dxdy$. Again $\mathbb{E}(h(X,Y))$ can be interpreted as population average or long run average of $h(X,Y)$.

Let $\mu_h = \mathbb{E}(h(X,Y))$, then $\mathrm{Var}(h(X,Y)) = \mathbb{E}[(h(X,Y) - \mu_h)^2]$.

Assume $(X,Y) \sim p(x,y)$. Consider $h(X,Y) = X+Y$. $\mathbb{E}(X+Y) = \sum_x \sum_y (x+y)p(x,y) = \sum_x \sum_y xp(x,y) + \sum_x \sum_y yp(x,y) = \mathbb{E}(X) + \mathbb{E}(Y)$. Similarly, assume $(X,Y) \sim f(x,y)$. Then $\mathbb{E}(X+Y) = \int\int (x+y)f(x,y)dxdy = \int\int xf(x,y)dxdy + \int\int yf(x,y)dxdy = \mathbb{E}(X) + \mathbb{E}(Y)$.

Let $\mu_X = E(X)$, $\mu_Y = E(Y)$, then $\mu_{X+Y} = \mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = \mu_X + \mu_Y$. Then

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}[((X+Y) - \mu_{X+Y})^2] \\
&= \mathbb{E}[((X-\mu_X) + (Y-\mu_Y))^2] \\
&= \mathbb{E}[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)] \\
&= \mathbb{E}[(X-\mu_X)^2] + \mathbb{E}[(Y-\mu_Y)^2] + 2\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y),
\end{aligned}$$

where $\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$ is called the covariance. It is defined for both discrete and continuous random variables.

A property of $\mathrm{Cov}(X,Y)$ is that $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. This is because $\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY - \mu_X Y - X\mu_Y - \mu_X\mu_Y] = \mathbb{E}(XY) - \mu_X\mathbb{E}(Y) - \mu_Y\mathbb{E}(X) + \mu_X\mu_Y = \mathbb{E}(XY) - \mu_X\mu_Y = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

Clearly, $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$ and $\mathrm{Cov}(Y,Y) = \mathrm{Var}(Y)$.

Geometric intuition using scatterplot. To understand $\mathrm{Cov}(X,Y)$ intuitively, we can imagine a population Ω of M people, and for each person Ω, let $X(\omega)$ be the height of Ω and $Y(\omega)$ be the weight of Ω. Then the M points $(X(\omega), Y(\omega))$ for $\omega \in \Omega$ form a cloud of points in the 2-dimensional domain. The density $f(x,y)$ tells us that the number of points in the cell $(x, x+\Delta x) \times (y, y+\Delta y)$ is $Mf(x,y)\Delta x\Delta y$. So for any function $h(X,Y)$,

$$\mathbb{E}(h(X,Y)) = \frac{1}{M}\sum_{s=1}^{M} h(X(\omega), Y(\omega)) = \frac{1}{M}\sum_{cells} h(x,y)Mf(x,y)\Delta x\Delta y \to \int\int h(x,y)f(x,y)dxdy.$$

That is, expectation is just population average. $\mathrm{Cov}(X,Y)$ is the population average of $(X(\omega) - \mu_X)(Y(s) - \mu_Y)$.

The vertical line $x = \mu_X$ and the horizontal line $y = \mu_Y$ divide the whole domain into four regions (small, big, skinny, chubby), which correspond to the four combinations of the signs of $X - \mu_X$ and $Y - \mu_Y$. If X and Y are positively related to each other, then there are more points falling into the regions where $(X-\mu_X)(Y-\mu_Y)$ are positive. So the overall covariance is positive. If X and Y are negatively related, then there are more points falling into the regions where $(X-\mu_X)(Y-\mu_Y)$ are negative, so the covariance is negative.

We can also change the population of points $\{(X(\omega), Y(\omega)), \omega \in \Omega\}$ into a large sample $\{(X_i, Y_i), i = 1, ..., n\}$ that are sampled from $f(x,y)$. Then $\mathbb{E}(h(X,Y))$ is the limit of

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i) = \frac{1}{n}\sum_{cells} h(x,y)nf(x,y)\Delta x\Delta y \to \int\int h(x,y)f(x,y)dxdy.$$

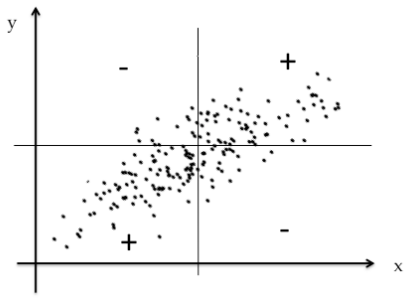

Figure 4: The sign of the covariance determines the trend of the linear relationship.

$\mathrm{Cov}(X,Y)$ measures the relationship between X_i and the corresponding Y_i for $i = 1,...,n$.

The following are the linear properties of covariance.

$$\begin{aligned}\mathrm{Cov}(aX+b,cY+d) &= \mathbb{E}[(aX+b-\mathbb{E}(aX+b))(cY+d-\mathbb{E}(cY+d))] \\ &= \mathbb{E}[a(X-\mathbb{E}(X))c(Y-\mathbb{E}(Y))] = ac\mathrm{Cov}(X,Y).\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(X+Y,Z) &= \mathbb{E}[(X+Y-\mathbb{E}(X+Y))(Z-\mathbb{E}(Z))] \\ &= \mathbb{E}[(X-\mathbb{E}(X)+Y-\mathbb{E}(Y))(Z-\mathbb{E}(Z))] \\ &= \mathbb{E}[(X-\mathbb{E}(X))(Z-\mathbb{E}(Z))]+\mathbb{E}[(Y-\mathbb{E}(Y))(Z-\mathbb{E}(Z))] \\ &= \mathrm{Cov}(X,Z)+\mathrm{Cov}(Y,Z).\end{aligned}$$

4.3 Correlation as cosine of angle

Covariance depends on units. For instance, if X is the height, and Y is the weight. If X is measured in meter and Y is measured in kilogram, then the unit of covariance is meter $\times$ kilogram. But if X is measured in inch and Y is measured in pound, then the covariance is in the unit of inch $\times$ pound.

We can standardize the variables by $X \to (X-\mu_X)/\sigma_X$, $Y \to (Y-\mu_Y)/\sigma_Y$. Then

$$\mathrm{Cov}\left(\frac{X-\mu_X}{\sigma_X},\frac{Y-\mu_Y}{\sigma_Y}\right) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \mathrm{Corr}(X,Y).$$

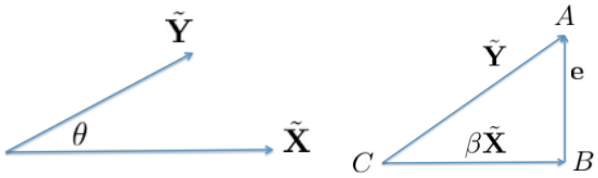

Figure 5: Correlation is cosine of an angle. The bigger the magnitude of the correlation, the smaller the magnitude of the error, and the stronger the linear relationship.

Geometric intuition using vector plot. Consider the population of M points $\{(X(\omega),Y(\omega)),\omega = 1,...,M\}$. Let $\mathbf{X} = (X(1),...,X(\omega),...,X(M))^\top$. Let $\mathbf{Y} = (Y(1),...,Y(\omega),...,Y(M))^\top$. Let $\tilde{\mathbf{X}} = (X(1)-\mu_X,...,X(\omega)-\mu_X,...,X(M)-\mu_X)^\top$. Let $\tilde{\mathbf{Y}} = (Y(1)-\mu_Y,...,Y(\omega)-\mu_Y,...,Y(M)-\mu_Y)^\top$. Then $|\tilde{\mathbf{X}}|^2 = M\mathrm{Var}(X)$. $|\tilde{\mathbf{Y}}|^2 = M\mathrm{Var}(Y)$. $\langle\tilde{\mathbf{X}},\tilde{\mathbf{Y}}\rangle = M\mathrm{Cov}(X,Y)$. So

$$\rho = \mathrm{Corr}(X,Y) = \frac{\langle\tilde{\mathbf{X}},\tilde{\mathbf{Y}}\rangle}{|\tilde{\mathbf{X}}||\tilde{\mathbf{Y}}|} = \cos\theta,$$

where θ is the angle between $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$.

We can also consider a large sample $(X_i, Y_i) \sim f(x,y)$, $i = 1,...,n$. Let $\bar{X} = \sum_{i=1}^n X_i/n \to \mathbb{E}(X)$, $\bar{Y} = \sum_{i=1}^n Y_i/n \to \mathbb{E}(Y)$. Let $\tilde{\mathbf{X}} = (X_1 - \bar{X}, ..., X_i - \bar{X}, ..., X_n - \bar{X})^\top$, and let $\tilde{\mathbf{Y}} = (Y_1 - \bar{Y}, ..., Y_i - \bar{Y}, ..., Y_n - \bar{Y})^\top$. Then for large n, $|\tilde{\mathbf{X}}|^2/n \to \text{Var}(X)$. $|\tilde{\mathbf{Y}}|^2/n \to \text{Var}(Y)$. $\langle \tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \rangle / n \to \text{Cov}(X,Y)$. So $\rho = \cos\theta$.

4.4 Correlation as the strength of regression

We can gain a deeper understanding of ρ from the regression perspective. In the vector plot, consider projecting $\tilde{\mathbf{Y}}$ on $\tilde{\mathbf{X}}$, and let $\beta\tilde{\mathbf{X}}$ be the projection. Let $\tilde{\mathbf{Y}} = \beta\tilde{\mathbf{X}} + \mathbf{e}$, where $\mathbf{e}$ is the residual vector that is perpendicular to $\tilde{\mathbf{X}}$. Then

$$\rho^2 = \cos\theta^2 = \frac{|\beta\tilde{\mathbf{X}}|^2}{|\tilde{\mathbf{Y}}|^2} = \beta^2 \frac{|\tilde{\mathbf{X}}|^2}{|\tilde{\mathbf{Y}}|^2} = 1 - \frac{|\mathbf{e}|^2}{|\tilde{\mathbf{Y}}|^2}$$

The bigger the ρ^2 is, the smaller $|\mathbf{e}|^2$ relative to $|\tilde{\mathbf{Y}}|^2$. Unpack the vectors, we have $Y_i - \bar{Y} = \beta(X_i - \bar{X}) + \varepsilon_i$, where $\mathbf{e} = (\varepsilon_i, i = 1,...,n)^\top$. This can be illustrated by the scatterplot, where the regression line is $y - \bar{Y} = \beta(x - \bar{X})$. If ρ^2 is large, then $\sum_{i=1}^n \varepsilon_i^2$ is small relative to $|\tilde{\mathbf{Y}}|^2/n \to \text{Var}(Y)$, so that the linear relationship is strong. Back to random variables,

$$Y - \mu_Y = \beta(X - \mu_X) + \varepsilon,$$

and

$$\rho^2 = \frac{\text{Var}(\beta X)}{\text{Var}(Y)} = \beta^2 \frac{\text{Var}(X)}{\text{Var}(Y)} = 1 - \frac{\text{Var}(\varepsilon)}{\text{Var}(Y)},$$

i.e., ρ^2 measures how much the variance of Y is explained by the regression on X. The above equation also tells us that $\beta = \rho\sigma_Y/\sigma_X$, and the regression line is

$$\frac{Y - \mu_Y}{\sigma_Y} = \rho \frac{X - \mu_X}{\sigma_Y},$$

that is, ρ is the regression coefficient or slope of the regression line for the standardized random variables. The initial meaning of "regression" means going back to the mean. The above relationship, with $|\rho| \le 1$, captures this meaning: the best prediction of Y is always closer to μ_Y than X is to μ_X.

4.5 Least squares derivation of regression

Suppose we want to predict Y by $\alpha + \beta X$. We can find α and β by minimizing $R(\alpha,\beta) = \mathbb{E}[(Y - \alpha - \beta X)^2]$. Let $\varepsilon = Y - \alpha - \beta X$.

$$\frac{\partial R(\alpha,\beta)}{\partial \alpha} = -2\mathbb{E}(\varepsilon) = 0.$$

So $\mathbb{E}(\varepsilon) = 0$. That is $\mathbb{E}(Y - \alpha - \beta X) = 0$, so $\mu_Y = \alpha + \beta\mu_X$. That is, the regression line $y = \alpha + \beta x$ goes through (μ_X, μ_Y). This can be easily understood from the scatterplot picture. Intuitively, $\mathbb{E}(\varepsilon) = 0$ means that some points are above the regression line while some points are below the regression line. Overall, the errors fluctuate around 0. If not, we can always update α to get a smaller $R(\alpha,\beta)$.

$$\frac{\partial R(\alpha,\beta)}{\partial \beta} = -2\mathbb{E}(\varepsilon X) = 0.$$

So $\text{Cov}(\varepsilon, X) = 0$. That is $\text{Cov}(Y - \alpha - \beta X, X) = 0$, i.e., $\text{Cov}(Y,X) = \beta\text{Cov}(X,X)$. Thus

$$\beta = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{\rho\sigma_X\sigma_Y}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}.$$

Intuitively, $\mathrm{Cov}(\varepsilon, X) = 0$ means that there is nothing in ε for X to explain. Otherwise, we can update β to get a smaller $R(\alpha, \beta)$. The variance of Y can be decomposed into the variance in regression and the variance in error:

$$\mathrm{Var}(Y) = \mathrm{Var}(\beta X + \varepsilon) = \mathrm{Var}(\beta X) + \mathrm{Var}(\varepsilon).$$

ρ^2 measures how much the variance in Y is explained by regression:

$$\frac{\mathrm{Var}(\beta X)}{\mathrm{Var}(Y)} = \frac{\beta^2 \mathrm{Var}(X)}{\mathrm{Var}(Y)} = \rho^2 \frac{\sigma_Y^2}{\sigma_X^2} \frac{\sigma_X^2}{\sigma_Y^2} = \rho^2.$$

4.6 Regression in terms of projections

Consider $\mathbf{1} = (1, ..., 1)^\top$ be a vector of n 1's. Its normalized version is $\mathbf{u} = \mathbf{1}/\sqrt{n}$. The projection of $\mathbf{X}$ on $\mathbf{1}$ is then $\langle \mathbf{X}, \mathbf{u} \rangle \mathbf{u} = \bar{X}\mathbf{1} \to \mathbb{E}(X)\mathbf{1}$. So projection on $\mathbf{1}$ is to take expectation. Thus $\tilde{\mathbf{X}} = \mathbf{X} - \bar{X}\mathbf{1}$ is the residual vector. Then according to Pythagorean, $|\mathbf{X}|^2 = |\bar{X}\mathbf{1}|^2 + |\tilde{\mathbf{X}}|^2$. So $\mathbb{E}(X^2) = \mathbb{E}(X)^2 + \mathrm{Var}(X)$. Also, $\bar{X}$ minimizes $\mathbb{E}((X-c)^2)$ among all possible c, because $|\mathbf{X} - c\mathbf{1}|^2 = |c\mathbf{1} - \mu_X \mathbf{1}|^2 + |\tilde{\mathbf{X}}|^2$, i.e., $\mathbb{E}((X-c)^2) = (c - \bar{X})^2 + \mathrm{Var}(X)$, which is minimized at $c = \bar{X} \to \mathbb{E}(X)$.

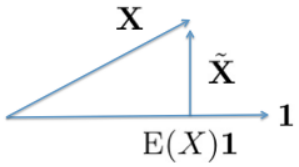

Figure 6: Geometry of centralization

Geometrically, we project $\mathbf{Y}$ on the subspace spanned by $\mathbf{1}$ and $\mathbf{X}$. The projection is $\alpha\mathbf{1} + \beta\mathbf{X}$. The residual $\mathbf{e} = \mathbf{Y} - \alpha\mathbf{1} - \beta\mathbf{X}$. The residual $\mathbf{e} \perp \mathbf{1}$. Thus $\mathbb{E}(\varepsilon) = 0$. Also $\mathbf{e} \perp \mathbf{X}$. So $\mathrm{Cov}(\varepsilon, X) = 0$.

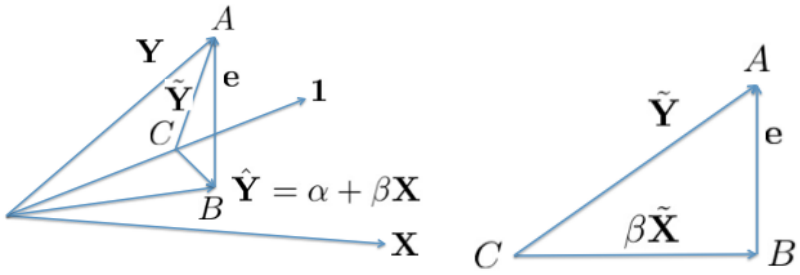

Figure 7: Geometry of linear regression

We can project both $\alpha\mathbf{1} + \beta\mathbf{X}$ and $\mathbf{Y}$ onto $\mathbf{1}$, and the projection meet at the same point: $(\alpha + \beta\bar{X})\mathbf{1} = \bar{Y}\mathbf{1}$, so $\bar{Y} = \alpha + \beta\bar{X}$. The vector from this point to $\alpha\mathbf{1} + \beta\mathbf{X}$ is $\beta\mathbf{X} - \beta\bar{X}\mathbf{1} = \beta(\mathbf{X} - \bar{X}\mathbf{1}) = \beta\tilde{\mathbf{X}}$. Consider the triangle formed by (C) this point, (A) $\tilde{\mathbf{Y}}$, which is the remainder of the projection of $\mathbf{Y}$ on $\mathbf{1}$, and (B) $\alpha\mathbf{1} + \beta\mathbf{X}$. By Pythagorean, $|\tilde{\mathbf{Y}}|^2 = |\beta\tilde{\mathbf{X}}|^2 + |\mathbf{e}|^2$, and $R^2 = |\beta\tilde{\mathbf{X}}|^2 / |\tilde{\mathbf{Y}}|^2 = \cos^2\theta = \rho^2$. R^2 is the proportion of variance in Y explained by the linear regression, and the magnitude of the correlation, $|\rho|$, measures the strength of the linear relationship.

4.7 Independence and uncorrelated

Two events A and B are independent if $P(A \cap B) = P(A)P(B)$.

An illustration of independence is that Ω is a unit square $[0,1]^2$. $A = [a_1, a_2] \times [0,1]$, and $B = [0,1] \times [b_1, b_2]$. Then $A \cap B = [a_1, a_2] \times [b_1, b_2]$. Then $P(A \cap B) = (a_2 - a_1)(b_2 - b_1) = P(A)P(B)$. A and B are perpendicular to each other, so we also write $A \perp B$ when A and B are independent.

For discrete random variables, if $p(x,y) = p_X(x)p_Y(y)$ for all (x,y), then X and Y are independent. In other words, the event $X = x$ and the event $Y = y$ are independent of each other for any (x,y). If X and Y are independent, then $P(Y = y|X = x) = P(Y = y)$, i.e., $p_{Y|X}(y|x) = p_Y(y)$ for any (x,y).

Similarly for continuous random variables, if $f(x,y) = f_X(x)f_Y(y)$ for all (x,y), then X and Y are independent. In other words, the event $X \in (x, x+\Delta x)$ and the event $Y \in (y, y+\Delta y)$ are independent for any (x,y).

If two random variables are independent, then they are uncorrelated, i.e., $\mathrm{Cov}(X,Y) = 0$. In the discrete case, if X and Y are independent, then

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)p(x,y) \\
&= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)p_X(x)p_Y(y) \\
&= \sum_x (x-\mu_X)p_X(x) \sum_y (y-\mu_Y)p_Y(y) \\
&= \left(\sum_x x p_X(x) - \mu_X\right)\left(\sum_y y p_Y(y) - \mu_Y\right) = 0.
\end{aligned}$$

The proof is similar for continuous random variables.

If $\mathrm{Cov}(X,Y) = 0$, however, they may not be independent. For instance, let X be a uniform distribution over $[-1,1]$. Let $Y = X^2$. Then X and Y are not independent. However, $\mathbb{E}(XY) = \mathbb{E}(X^3) = 0$, and $\mathbb{E}(X) = 0$. Thus $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$.

4.8 Multivariate statistics

Whenever possible, you can always think of matrices or vectors as scalers or numbers in your calculations. You only need to take care to match the dimensionalities. We shall define expectations, variances, and derivatives for vectors as a matter of conveniently packaging their elements so that we can avoid subindices in our calculations.

Expectation of a random matrix. Consider a random matrix X. Suppose X is $m \times n$, and the elements of X are x_{ij}, $i = 1,...,m$ and $j = 1,...,n$. Usually we write $X = (x_{ij})_{m\times n}$ or simply $X = (x_{ij})$. We define

$$\mathbb{E}(X) = (\mathbb{E}(x_{ij})),$$

i.e., taking expectations element-wise. Let A be a constant matrix of appropriate dimension, then $\mathbb{E}(AX) = A\mathbb{E}(X)$. Let B be another constant matrix of appropriate dimension, then $\mathbb{E}(XB) = \mathbb{E}(X)B$. The proof follows the linear property of expectation. Let $Y = AX$, then $y_{ij} = \sum_k a_{ik}x_{kj}$, and $\mathbb{E}(y_{ij}) = \mathbb{E}(\sum_k a_{ik}x_{kj}) = \sum_k a_{ik}\mathbb{E}(x_{kj})$. Thus $\mathbb{E}(Y) = (\mathbb{E}(y_{ij})) = (\sum_k a_{ik}\mathbb{E}(x_{kj})) = A\mathbb{E}(X)$.

The above result can be easily understood if we have iid copies $X_1,...,X_n$, so that $\sum_{i=1}^n X_i/n \to \mathbb{E}(X)$, and $\sum_{i=1}^n AX_i/n \to \mathbb{E}(AX)$, but $\sum_{i=1}^n AX_i/n = A\sum_{i=1}^n X_i/n \to A\mathbb{E}(X)$. Thus $\mathbb{E}(AX) = A\mathbb{E}(X)$.

Variance-covariance matrix of a random vector. Let X be a random vector. Let $\mu_X = \mathbb{E}(X)$. We define

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu_X)(X-\mu_X)^\top].$$

Then the (i,j)-th element of $\mathrm{Var}(X)$ is $\mathrm{Cov}(x_i,x_j)$. Let A be a constant matrix of appropriate dimension, then $\mathrm{Var}(AX) = A\mathrm{Var}(X)A^\top$. This is because

$$\begin{aligned}
\mathrm{Var}(AX) &= \mathbb{E}[(AX - \mathbb{E}(AX))(AX - \mathbb{E}(AX))^\top] \\
&= \mathbb{E}[(AX - A\mu_X)(AX - A\mu_X)^\top] \\
&= \mathbb{E}[A(X-\mu_X)(X-\mu_X)^\top A^\top] \\
&= A\mathbb{E}[(X-\mu_X)(X-\mu_X)^\top]A^\top \\
&= A\mathrm{Var}(X)A^\top.
\end{aligned}$$

We can also define $\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)^\top]$, then

$$\begin{aligned}
\mathrm{Cov}(AX,BY) &= \mathbb{E}[(AX - A\mu_X)(BY - B\mu_Y)^\top] = \mathbb{E}[A(X-\mu_X)(Y-\mu_Y)^\top B^\top] \\
&= A\mathbb{E}[(X-\mu_X)(Y-\mu_Y)^\top]B^T = A\mathrm{Cov}(X,Y)B^\top
\end{aligned}$$

4.9 Multivariate normal

Let $X \sim \mathrm{N}(\mu,\Sigma)$. The density of multivariate normal or multivariate Gaussian is

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

The maximum of $\log f(x)$ is μ, which is the solution to $d\log f(x)/dx = 0$. The second derivative matrix $d^2\log f(x)/dx^2 = (\partial^2 \log f(x)/\partial x_i \partial x_j)$ is Σ^{-1}. So if the log of the density function $\log f(x)$ is quadratic, we can identify μ and Σ by the first and second derivatives of $\log f(x)$.

For simplicity, we shall assume that $\mu = 0$ in the following. Otherwise we can simply let $X' = X - \mu$ to centralize X.

4.10 Eigen decomposition and principal component analysis

Σ can be decomposed into $\Sigma = Q\Lambda Q^\top$, where $Q = (q_1, q_2 \ldots q_n)$ are orthonormal vectors, i.e., $\langle q_i, q_j\rangle = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ is a diagonal matrix. Q forms a set of basis. We can expand X in Q, so that $X = \sum_i q_i z_i = QZ$, and z_i are the coordinates of X in the basis Q, i.e., $z_i = \langle X, q_i\rangle$, so $Z = Q^\top X$. Thus $Q^\top Q = QQ^\top = I$. $|\Sigma| = |\Lambda|$. The density of Z is

$$\begin{aligned}
f(z) &= \frac{1}{(2\pi)^{n/2}|\Lambda|^{1/2}} \exp\left(-\frac{1}{2}(QZ)^\top \Sigma^{-1} QZ\right) \\
&= \frac{1}{(2\pi)^{n/2}|\Lambda|^{1/2}} \exp\left(-\frac{1}{2}Z^\top \Lambda^{-1} Z\right) \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left(-\frac{z_i^2}{2\lambda_i}\right),
\end{aligned}$$

where the Jacobian of $X = QZ$ is 1 since the transformation is just a rotation. Thus $z_i \sim \mathrm{N}(0,\lambda_i)$ independently. So $\mathbb{E}(Z) = 0$ and $\mathrm{Var}(Z) = \Lambda$. As a result $\mathbb{E}(X) = 0$ and $\mathrm{Var}(X) = Q\Lambda Q^\top = \Sigma$.

The eigen decomposition enables us to view $\mathrm{N}(0,\Sigma)$ in a simple way. We rotate the basis so that they are aligned with the axes of the elliptical contours of $f(x)$.

Assuming $\lambda_1 > \lambda_2 > ... > \lambda_n$. If $\lambda_i = \mathrm{Var}(z_i)$ is very small for $i > m$, then $z_i \approx 0$ for $i > m$ (recall $\mathbb{E}(z_i) = 0$). We can represent $X \approx \sum_{i=1}^m q_i z_i$, thus reducing the dimensionality of X from n to m. The $(q_i, i = 1, ..., m)$ are called principal components. For instance, if X is a face image, then $(q_i, i = 1, ..., m)$ are the eigen faces, which may correspond to different features of a face (e.g., eyes, nose, mouth etc.), and $(z_i, i = 1, ..., m)$ is a low dimensional indexing of X.

5 Conditional distribution and expectation

5.1 Conditional probability

In Example 1, let A be the event that the number is 6, and let B be the event that the number is greater than 4. Then $P(A) = 1/6$, but $P(A|B) = 1/2$. It is as if we randomly sample from $B = \{5,6\}$, as if B is the new sample space.

In Example 2, let A be the event that the height is taller than 6 feet, B be the event that the gender is male. Then $P(A)$ is the proportion of tall people in the whole population, and $P(A|B)$ is the proportion of tall people within the male sub-population,

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

It is as if we randomly sample a person from the male sub-population B, as if B is the new sample space Ω, and $A \cap B$ is the new A. $P(B|A)$ is the proportion of males among the tall people.

In Example 3, A and B are two regions. In calculating $P(A|B)$, it is as if we randomly throw the point into B, as if B is the new sample space Ω, $P(A|B)$ is the probability of falling into $A \cap B$, as if $A \cap B$ is the new A, so again we have

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

In general, we can define

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

This can be considered the fourth axiom of probability.

From long run frequency perspective, $P(A)$ tells us how often A happens in the long run. $P(A|B)$ tells us how often A happens when B happens. Suppose we repeat the experiments 10,000 times independently. If B happens $3,000$ times. A and B happen together $1,000$ times. Then when B happens, A happens 1/3 of times. Thus $P(A|B) = 1/3 = P(A \cap B)/P(B) = (1000/10000)/(3000/10000)$.

$P(A|B)$ and $P(B|A)$ are in general not the same. For example, let A be the event that there is fire alarm, and B be the event that there is fire. Then $P(A|B)$ is close to 1, but $P(B|A)$ is close to 0.

5.2 Conditional probability behaves like regular probability

The conditional probability is really a regular probability under a new re-imagined experiment. As long as the condition is fixed, then we are talking about the same re-imagined experiment, and all the conditional probabilities under this same condition just behave like regular probabilities in this re-imagined experiment, i.e., $P(A|B)$ behaves like a regular probability in that it satisfies the three axioms of probability as long as B is fixed:

- For any event A, $P(A|B) \geq 0$.
- For the sample space Ω, $P(\Omega|B) = 1$.
- If $A_1 \cap A_2 = \phi$, then $P(A_1 \cup A_1|B) = P(A_1|B) + P(A_2|B)$.

You can prove the above three properties easily using the definition of conditional probability. As a result, for any rule of probability calculation, there is a corresponding rule for the conditional probability, as long as the condition is fixed.

For instance, (1) $P(A) = 1 - P(A^c) \to P(A|B) = 1 - P(A^c|B)$. (2) $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \to P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$.

In fact, from $P(A|B) = P(A \cap B)/P(B)$, we have $P(A|B \cap C) = P(A \cap B|C)/P(B|C)$, where we insert to both sides the common condition C.

The meta rule is that for whatever property of probability, either in the form of an equation or an inequality, as long as we insert or add the same condition into all the probabilities, the equation or inequality still holds. Of course, whatever results from the meta rule can be rigorously proved.

While conditional probability behaves like regular probability, regular probability is actually also the conditional probability, $P(A) = P(A|\Omega)$, where the condition is the sample space and is often made implicit.

Sometimes, the condition actually defines a physically new experiment. For instance, consider a random walk on the integers. Let X_t be the position at time t. Then $P(X_{t+1} = 4|X_t = 5) = 1/2 = P(X_{t+1} = 6|X_t = 5)$. We do not have to use the definition $P(A|B) = P(A \cap B)/P(B)$ to calculate it, as $P(A|B)$ is given. In fact, in this case, we can calculate $P(A \cap B) = P(B)P(A|B)$, which is the chain rule.

5.3 Conditional distribution

By the definition of conditional probability, $P(Y = y|X = x) = P(X = x, Y = y)/P(X = x)$. We denote $P(Y = y|X = x)$ by $p_{Y|X}(y|x)$. Then $p_{Y|X}(y|x) = p(x,y)/p_X(x)$. Here $p_{Y|X}$ is a single notation. We call $p_{Y|X}(y|x)$ the conditional distribution. Similarly, $p_{X|Y}(x|y) = p(x,y)/p_Y(y)$.

For population interpretation, $p_{Y|X}(y|x)$ tells us: among all the people in the sub-population $\{\omega : X(\omega) = x\}$, what is the fraction of those with $Y(\omega) = y$.

As to long run frequency, $p_{Y|X}(y|x)$ tells us: when $X = x$, how often $Y = y$. This conditional probability can be used to predict Y based on X.

If both X and Y are continuous,

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{P(Y \in (y, y+\Delta y)|X \in (x, x+\Delta x))}{\Delta y} \\ &= \frac{P(Y \in (y, y+\Delta y), X \in (x, x+\Delta x))}{P(X \in (x, x+\Delta x))\Delta y} \\ &= \frac{f(x,y)\Delta x \Delta y}{f_X(x)\Delta x \Delta y} \\ &= \frac{f(x,y)}{f_X(x)}. \end{aligned}$$

For the population Ω, we can think of $(X(\omega), Y(\omega))$ as the coordinate of a point. So if there are M people in Ω, then we have M points. If M is large, then these M points are like a cloud. $f_{Y|X}(y|x)$ tells us the distribution of the points in the vertical slice of $(x, x+\Delta x)$. $f_{X|Y}(x|y)$ tells us the distribution of the points in the horizontal slice of $(y, y+\Delta y)$. For instance $f_{Y|X}(y|x = 6ft)$ is the distribution of weights among people who are 6 ft tall. Clearly this is different from the distribution of weights among people who are 5ft tall.

The above interpretation can also be applied to a large sample $(X_i, Y_i) \sim f(x,y)$ independently for $i = 1, ..., n$. The scatterplot of $\{(X_i, Y_i), 1, ..., n\}$ form a cloud, whose density is described by $f(x,y)$. $f_{Y|X}(y|x)\Delta y$ tells us when $X_i \in (x, x+\Delta x)$, how often $Y_i \in (y, y+\Delta y)$, i.e., the vertical slice of the scatterplot for $(x, x+\Delta x)$.

5.4 Conditional distribution of multivariate normal

Let us partition X into X_1, X_2.

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Let $\varepsilon = X_2 - AX_1$. We choose A to make $\mathrm{Cov}(\varepsilon, X_1) = 0$,

$$\begin{aligned}\mathrm{Cov}(\varepsilon, X_1) &= \mathrm{Cov}(X_2 - AX_1, X_1) \\ &= \mathrm{Cov}(X_2, X_1) - A\mathrm{Cov}(X_1, X_1) \\ &= \Sigma_{21} - A\Sigma_{11} = 0,\end{aligned}$$

so $A = \Sigma_{21}\Sigma_{11}^{-1}$, and $X_2 = AX_1 + \varepsilon$. This can be considered a regression of X_2 on X_1. The residual variance is

$$\begin{aligned}\mathrm{Var}(\varepsilon) &= \mathrm{Cov}(\varepsilon, \varepsilon) \\ &= \mathrm{Cov}(X_2 - AX_1, \varepsilon) \\ &= \mathrm{Cov}(X_2, \varepsilon) \\ &= \mathrm{Cov}(X_2, X_2 - AX_1) \\ &= \mathrm{Cov}(X_2, X_2) - \mathrm{Cov}(X_2, AX_1) \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.\end{aligned}$$

Thus

$$\begin{pmatrix} X_1 \\ \varepsilon \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix} \right).$$

So the marginal distribution $X_1 \sim \mathrm{N}(0, \Sigma_{11})$, and the conditional distribution

$$[X_2|X_1] \sim \mathrm{N}(\Sigma_{21}\Sigma_{11}^{-1}X_1, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}).$$

5.5 Conditional expectation and variance

The linear relationship may not give us the best prediction. Given $X = x$, $Y \sim f_{Y|X}(y|x)$, which is a slice or stratum of the population, for instance, people of 40 years old if $x = 40$. According to our discussion above, for any random variable Y, the best constant for predicting Y is μ_Y. Given $X = x$, the best prediction of Y is

$$\mathbb{E}(Y|X = x) = \int y f_{Y|X}(y|x)dy = h(x).$$

Since $h(x)$ is the best prediction of Y for each $X = x$, $h(X)$ is the best prediction of Y overall.

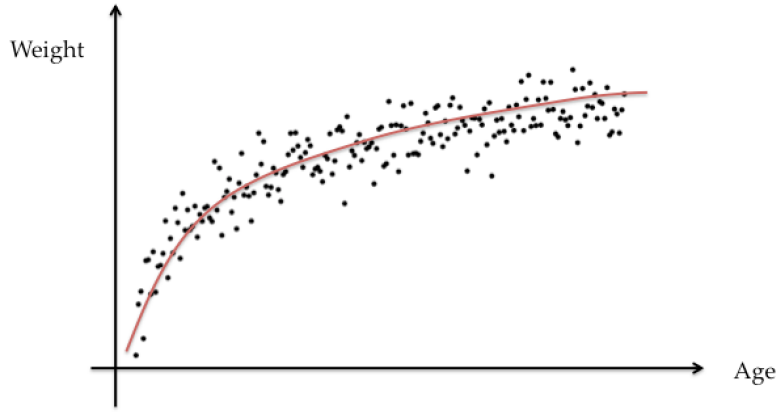

Figure 8: Regression curve $h(x) = \mathbb{E}(Y|X = x)$, e.g., average weight of 40 years old is $h(40)$.

In the scatterplot picture, $h(x)$ is the regression curve. For instance, if Y is weight, and X is age, then $h(x)$ tells us the average weight of 40 years old if $x = 40$.

Treat $h(x)$ as a function or transformation, we can write $h(X) = \mathbb{E}(Y|X)$.

$$\mathbb{E}(h(X)) = \int h(x) f_X(x) dx = \int \int y f_{Y|X}(y|x) dy f_X(x) dx = \int \int y f(x,y) dx dy = \mathbb{E}(Y).$$

So we have the Adam formula

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)).$$

Intuitively, we can imagine changing each person's weight $Y(\omega)$ to its group average $h(X(\omega))$. This change won't change the total weight of each group. Thus it won't change the total weight for the whole population, so it won't change the average for the whole population, i.e., $\mathbb{E}(h(X)) = \mathbb{E}(Y)$.

We can also define the conditional variance as

$$\text{Var}(Y|X = x) = \mathbb{E}[(Y - h(X))^2 | X = x] = \int (y - h(x))^2 f_{Y|X}(y|x) dy,$$

which is the variance of weights of 40 years old if $x = 40$.

Let $\varepsilon = Y - h(X)$, then $Y = h(X) + \varepsilon$. $\mathbb{E}(\varepsilon) = \mathbb{E}(Y - h(X)) = \mathbb{E}(\mathbb{E}(Y - h(X)|X)) = 0$. $\mathbb{E}(\varepsilon g(X)) = \mathbb{E}(\mathbb{E}(\varepsilon g(X)|X)) = 0$. So $\text{Cov}(\varepsilon, g(X)) = 0$ for any function $g(x)$. This generalizes the results of linear regression, where $h(x)$ and $g(x)$ are restricted to be linear functions.

Then

$$\begin{aligned}
\text{Var}(Y) &= \text{Var}(h(X) + \varepsilon) \\
&= \text{Var}(h(X)) + \text{Var}(\varepsilon) \\
&= \text{Var}(\mathbb{E}(Y|X)) + \mathbb{E}((Y - h(X))^2) \\
&= \text{Var}(\mathbb{E}(Y|X)) + \mathbb{E}(\mathbb{E}((Y - h(X))^2 | X)) \\
&= \text{Var}(\mathbb{E}(Y|X)) + \mathbb{E}(\text{Var}(Y|X)).
\end{aligned}$$

This is the Eve formula, which says that the total variance can be decomposed into the between group variance and the within group variance. This can be seen clearly from the scatterplot picture. Take things to the extreme, if $Y = h(X)$ exactly, then the within group variance is zero. So the variance of Y is the same as the between group variance. If Y is independent of X, then $\mathbb{E}(Y|X) = \mu_Y$, so the between group variance is zero. So the variance of Y is the same as the within group variance.

In the vector picture, $h(\mathbf{X})$ is the projection of $\mathbf{X}$ onto the space of all functions of $\mathbf{X}$ (for a function $g(x)$, $g(\mathbf{X})$ means the vector $(g(X_1), ..., g(X_i), ..., g(X_n))^\top$. Let $\mathbf{e} = \mathbf{Y} - h(\mathbf{X})$, then $\mathbf{e} \perp g(\mathbf{X})$ for any $g(\mathbf{X})$, including $\mathbf{1}$. From $\mathbf{e} \perp \mathbf{1}$, we have $\mathbb{E}(\varepsilon) = 0$. From $\mathbf{e} \perp g(\mathbf{X})$, we have $\mathbb{E}(\varepsilon g(X)) = \text{Cov}(\varepsilon, g(X)) = 0$.

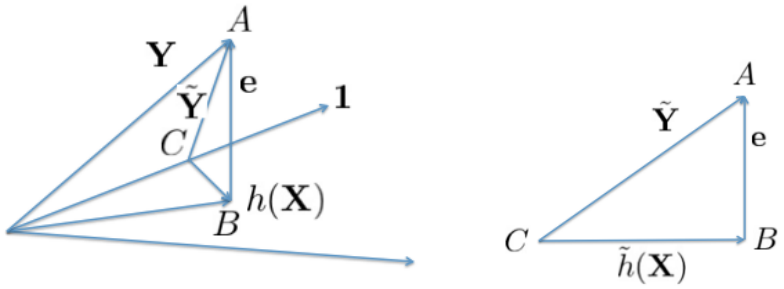

Figure 9: Geometry of Adam and Eve

If we project both $h(\mathbf{X})$ and $\mathbf{Y}$ onto $\mathbf{1}$, they meet at the same point, which is $\mathbb{E}(h(X))\mathbf{1}$ or $\mathbb{E}(Y)\mathbf{1}$. Thus $\mathbb{E}(h(X)) = \mathbb{E}(Y)$.

Consider the triangle formed by $\mathbf{Y}$, $h(\mathbf{X})$, and $\mathbb{E}(Y)\mathbf{1}$, Pythagorean says $|\tilde{\mathbf{Y}}|^2 = |\mathbf{e}|^2 + |h(\tilde{\mathbf{X}})|^2$, thus $\text{Var}(Y) = \mathbb{E}(\text{Var}(Y|X)) + \text{Var}(\mathbb{E}(Y|X))$.

So the same picture contains both the Adam formula and the Eve formula.

We can also define $R^2 = \text{Var}(\mathbb{E}(Y|X))/\text{Var}(Y)$ in analogy to linear regression.

5.6 Conditional covariance

A generalization of the Eve formula involves the conditional covariance. For three random variables X, Y, Z, let $h(Z) = \mathbb{E}(X|Z)$, $g(Z) = \mathbb{E}(Y|Z)$. Define

$$\begin{aligned}\text{Cov}(X,Y|Z=z) &= \mathbb{E}[(X-\mathbb{E}(X|Z))(Y-\mathbb{E}(Y|Z))|Z=z] \\ &= \mathbb{E}[(X-h(Z))(Y-g(Z))|Z=z] \\ &= \int\int (x-h(z))(y-g(z))f_{X,Y|Z}(x,y|z)dxdy.\end{aligned}$$

Let $\varepsilon = X - h(Z)$, and $\delta = Y - g(Z)$, then

$$\begin{aligned}\text{Cov}(X,Y) &= \text{Cov}(h(Z)+\varepsilon, g(Z)+\delta) \\ &= \text{Cov}(h(Z),g(Z)) + \text{Cov}(\varepsilon,\delta) \\ &= \text{Cov}(\mathbb{E}(X|Z),\mathbb{E}(Y|Z)) + \mathbb{E}(\mathbb{E}((X-h(Z))(Y-g(Z))|Z)) \\ &= \text{Cov}(\mathbb{E}(X|Z),\mathbb{E}(Y|Z)) + \mathbb{E}(\text{Cov}(X,Y|Z)).\end{aligned}$$

It is possible that $\text{Cov}(X,Y|Z) \leq 0$, but $\text{Cov}(X,Y) > 0$, or the other way around. For instance, let X be SAT score and Y be the GPA of those students admitted into a school. It is possible that within a school, the covariance between X and Y is usually negative. But if we look across all the schools, X and Y are positively correlated.

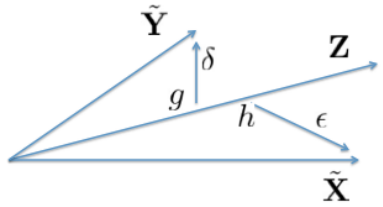

Figure 10: Geometry of conditional covariance

In the vector picture, consider the vectors $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$. If we project them onto the space of all functions of $\mathbf{Z}$, and let ε and δ be the remainders. Then it is possible that the angle between ε and δ is greater than $\pi/2$, but the angle between $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ is less than $\pi/2$. You can also change the direction of $\tilde{\mathbf{Y}}$ to make it the other way around.

5.7 Chain rule and rule of total probability

Chain rule: $P(A\cap B) = P(B)P(A|B)$. This rule can be generalized to $P(A_1\cap A_2\cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_2\cap A_1)$. Sometimes we simply write $P(A_3|A_2\cap A_1) = P(A_3|A_2,A_1)$.

Rule of total probability: Suppose $C_1,...,C_n$ partition Ω, i.e., $C_i\cap C_j = \phi$ for $i\neq j$, and $\cup_{i=1}^n C_i = S$. Then $P(A) = P(\cup_{i=1}^n (C_i\cap A)) = \sum_{i=1}^n P(C_i\cap A) = \sum_{i=1}^n P(C_i)P(A|C_i)$. This is a sum-product rule, where we sum over the alternative chains of events. $C_1,...,C_n$ can be considered an exhaustive list of mutually exclusive causes, and A is the effect.

5.8 Conditional independence

Following the meta rule about conditional probability, we can define conditional independence by inserting a fixed condition. We say that A and B are conditionally independent given C, $A \perp B|C$, if $P(A|B \cap C) = P(A|C)$, which is the Markov property, or $P(A \cap B|C) = P(A|C)P(B|C)$, which is the shared cause property.

We can also express independence or conditional independence by random variables, where A, B, and C are basic events, such as $X = x$, $Y = y$ and $Z = z$ in the discrete case, or $X \in (x, x+\Delta x)$, $Y \in (y, y+\Delta y)$ and $Z \in (z, z+\Delta z)$ in the continuous case.

We say $X \perp Y$ if $p(x,y) = p_X(x)p_Y(y)$ $\forall(x,y)$. For continuous case, it will be $f(x,y) = f_X(x)f_Y(y)$ $\forall(x,y)$. You can think of $f(x,y)$ as describing the density of a cloud of points $(X_i, Y_i) \sim f(x,y)$ independently for $i = 1, ..., n$. This cloud of points show no trend or pattern. For instance, if both X and Y are standard normal random variables, then the cloud is a blob of points.

We say $X \perp Y|Z$ if $p_{XY|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$, which is the shared cause property, illustrated by a causal diagram with Z pointing to both X and Y. We can also write $p_{Y|XZ}(y|x,z) = p_{Y|Z}(y|z)$, which is the Markov property, illustrated by a causal diagram $X \to Z \to Y$.

For continuous case, the shared cause property is $f(x,y|z) = f(x|z)f(y|z)$. Consider the model $X = Z + \varepsilon_1$ and $Y = Z + \varepsilon_2$, where ε_1 and ε_2 are independent standard normal random variables, and they are independent of Z that follows a certain distribution. In terms of the cloud of points, each $f(x,y|z)$ is a blob of points centered at (z,z), and there is no trend or pattern. However, if merge all the blobs of points centered at (z,z) across different z, the resulting cloud of points is $f(x,y)$. Then clearly there is a trend in $f(x,y)$. That is, X and Y are not independent marginally. More generally, we can encode the shared cause property $X \perp Y|Z$ by the structural equations $Z = h_Z(\varepsilon_Z)$, $X = h_X(Z, \varepsilon_X)$ and $Y = h_Y(Z, \varepsilon_Y)$, where all the ε's are independent.

For continuous case, the Markov property is $f(y|x,z) = f(y|z)$. Consider the model $Z = X + \varepsilon_1$ and $Y = Z + \varepsilon_2$, where ε_1 and ε_2 are independent standard normal random variables, and they are independent of Z that follows a certain distribution. Then given X and Z, $Y \sim \mathrm{N}(Z, 1)$, which has nothing to do with X. But if we do not know Z and we only know X, then $Y \sim \mathrm{N}(X, 2)$. Intuitively, we can think of Z as a cloud of points around X. Each $[Y|Z = z]$ is a cloud of points around z. If we merge all the clouds of points across different Z, we get a cloud of points around X. This can also be understood in terms of quincunx or Galton's board. More generally, we can encode the Markov property $X \to Z \to Y$ by $X = h_X(\varepsilon_X)$, $Z = h_Z(X, \varepsilon_Z)$ and $Y = h_Y(Z, \varepsilon_Y)$, where all the ε's are independent.

5.9 Markov property

Suppose we make random walk on three states $S = \{1, 2, 3\}$. At each step, we randomly move to one of the other two states, regardless of the past history. Let X_t be the state at step t. Then $P(X_{t+1} = y|X_t = x, X_{t-1}, X_{t-2}, ..., X_0) = P(X_{t+1} = y|X_t = x) = K(x,y)$. That is, knowing the current state x, the probability that we end up in state y in the next step is determined, regardless where we come from.

A very important property of the Markov chain is $P(X_{t+1} = y) = \sum_x P(X_{t+1} = y, X_t = x) = \sum_x P(X_t = x)P(X_{t+1} = y|X_t = x) = \sum_x P(X_t = x)K(x,y)$. This is a direct application of the rule of total probability. $P(X_t = x)$ is called the marginal probability.

An intuitive way of understanding the above result is as follows. Consider a population of 1 million people move around the three states according to the random walk described above. We can interpret $P(X_t = x) \times 1$ *million* as the number of people in state x at time t. $P(X_{t+1} = y|X_t = x) = K(x,y)$ is the fraction of those in state x who will move to state y. Then $P(X_t = x)K(x,y)$ is the number of people who will move from x to y (i.e., those people who are in state x at time t and who will be in state y at time $t+1$). Thus $\sum_x P(X_t = x)K(x,y)$ will be the total number of people who will end up in y at time $t+1$, and this is $P(X_{t+1} = y) \times 1$ *million*. So you can visualize the change of the probability distribution of $(P(X_t = x), \forall x \in S)$

as the change of the distribution of the population over time t, as if we are watching a movie where each t is a frame of the movie. As $t \to \infty$, the distribution (either a probability distribution or a population distribution) will converge to a stationary distribution π, i.e., $P(X_t = x) \to \pi(x)$ for all $x \in S$, no matter what is the initial distribution or no matter where we start from.

In the above example, the conditional independence assumption enables us to conclude that among all the people in x at time t, the fraction of people who will move to y at time $t+1$ is $K(x,y)$ regardless of these people's past immigration histories.

The google pagerank algorithm assumes a random walk or Markov chain over the state space of all the webpages. The stationary distribution $\pi(x)$ can be interpreted as the proportion of web surfers who are on page x. So $\pi(x)$ is an indicator of the popularity of x.

5.10 Bayes rule

Bayes rule: Given the prior probability $P(C_i)$, and the causal conditional probability $P(A|C_i)$, for $i = 1, ..., n$, we can get the posterior probability $P(C_i|A)$ by

$$\begin{aligned} P(C_i|A) &= \frac{P(C_i \cap A)}{P(A)} \\ &= \frac{P(C_i \cap A)}{\sum_{j=1}^{n} P(C_j \cap A)} \\ &= \frac{P(C_i)P(A|C_i)}{\sum_{j=1}^{n} P(C_j)P(A|C_j)}, \end{aligned}$$

where we use j in the denominator just to avoid confusion with i in the numerator.

You can think about the rule of total probability and the Bayes rule in terms of the Venn diagram. $P(C_i)$ is the area of C_i. $P(A|C_i)$ is the proportion of C_i occupied by A. $P(C_i|A)$ is the proportion of A occupied by C_i.

As in most of the probability calculations, the above rules can be reduced to elementary school problems of counting. For instance, let C_1 be the event that a random person is male, and C_2 be the event that the random person is female. Then our prior belief is that $P(C_1) = P(C_2) = 1/2$. Let A be the event that this person is taller than 6ft. Suppose $P(A|C_1) = 10\%$, i.e., 10% of females are taller than 6ft. $P(A|C_2) = 50\%$, i.e., 50% of males are taller than 6ft. Then according to the rule of total probability, $P(A) = P(C_1)P(A|C_1) + P(C_2)P(A|C_2) = 1/2 \times 10\% + 1/2 \times 50\% = 30\%$. That is, overall, 30% of the whole population are taller than 6ft. If a person is taller than 6ft, then the chance that this person is male is $P(C_1|A) = P(C_1 \cap A)/P(A) = P(C_1)P(A|C_1)/P(A) = 25\%/30\% = 5/6$, i.e., among people taller than 6ft, 5/6 of them are males.

The above is 5th grade problem. You can make it a 3rd grade problem by counting a population of 100 people, with 50 males and 50 females. Among the 50 males, 25 are taller than 6ft. Among the 50 females, 5 of them are taller than 6ft. So among all the 100 people, 30 of them taller than 6ft. Among these 30 people who are taller than 6ft, 25 of them are males. So the proportion is 5/6.

In terms of random variables, suppose we randomly sample a person from the population, for a person $\omega \in \Omega$, let $X(\omega) \in \{male, female\}$ be the gender of Ω, and let $Y(s) \in \{short, medium, tall\}$ be the height of Ω (say, short means the person is shorter than 5ft, tall means that the person is taller than 6ft, and medium means in between). Let $C_1 = \{s : X(\omega) = male\}$, and $C_2 = \{\omega : X(\omega) = female\}$. Let $A = \{\omega : Y(\omega) = tall\}$. For simplicity, we can just write C_1 as $X = male$, and C_2 as $X = female$, and A as $Y = tall$. Then we can write the rule of total probability $P(Y = y) = \sum_x P(X = x, Y = y) = \sum_x P(X = x)P(Y = y|X = x)$, where

$x \in \{male, female\}$, and $y \in \{short, medium, tall\}$. We can write the Bayes rule as

$$\begin{aligned} P(X=x|Y=y) &= \frac{P(X=x,Y=y)}{P(Y=y)} \\ &= \frac{P(X=x,Y=y)}{\sum_{x'} P(X=x',Y=y)} \\ &= \frac{P(X=x)P(Y=y|X=x)}{\sum_{x'} P(X=x')P(Y=y|X=x')}, \end{aligned}$$

where we use x' in the denominator just to avoid confusion with x in the numerator. Again, $x, x' \in \{male, female\}$, and $y \in \{short, medium, tall\}$.

For the rest of the course, we shall usually use random variables notation instead of events notation, but they can be translated into each other. It is just a matter of language.

The prior probability $P(X=x)$ and the generative/causal conditional probability $P(Y=y|X=x)$ is in the direction $X \to Y$. The Bayes rule is in the inverse/inference/diagnosis direction $Y \to X$. The former is like setting up an equation, and the latter is like solving for the unknown X from Y. The Bayes rule provides us a consistent way of making inference for cause-effect analysis, such as disease-symptom analysis.

The Bayes rule can serve as a basis for reasoning and inference, such as inverse problem, cause-effect analysis, and disease-symptom diagnosis.

For Bayes rule, let the prior distribution be $X \sim p_X(x)$. Let the conditional distribution be $[Y|X = x] \sim P_{Y|X}(y|x)$. Then the joint distribution is $p(x,y) = p_X(x)p_{Y|X}(y|x)$. The marginal distribution of Y is $p_Y(y) = \sum_x p(x,y)$. The posterior distribution is $[X|Y=y] \sim p_{X|Y}(x|y) = p(x,y)/p(y)$. For continuous random variable, the probability mass function becomes probability density function and summation becomes integral.

$[Y|X]$ is the generative or causal direction (from cause to effect, or from disease to symptom, or setting up an equation). $[X|Y]$ is the inferential or inference direction (infer cause from effect, diagnosis of disease from symptom, or finding the solution to the equation).

The following are 4 examples. One can derive the results using the Bayes rule with the basic events, and interpret the results intuitively by counting the populations or drawing graphs. The following is a formal treatment using the above formula.

5.11 Fire alarm example

Both X and Y are discrete. Let $X=1$ denote that there is fire, and $Y=1$ denote that there is alarm. Let $P(X=1) = p_X(1) = \rho, P(Y=1|X=1) = p_{Y|X}(1|1) = \alpha, P(Y=1|X=0) = p_{Y|X}(1|0) = \beta$. Then

$$p_Y(1) = \sum_x p(x,1) = \sum_x p_X(x)p_{Y|X}(1|x) = \rho\alpha + (1-\rho)\beta.$$

$$p_{X|Y}(1|1) = \frac{p(1,1)}{P_Y(1)} = \frac{p_X(1)P_{Y|X}(1|1)}{p_Y(1)} = \frac{\rho\alpha}{\rho\alpha + (1-\rho)\beta}.$$

The above calculation also underlies the rare disease example. Suppose 1% of population is inflicted with a rare disease. Suppose a random person gets a test. The accuracy of the test is such that if the person has disease, the probability of getting a positive result is 90%. If the person does not have the disease, the probability of getting a negative result is 90%. We want to calculate the probability that the person has disease given the fact that the person is tested positive. Following the above calculation, $P(disease|+) = 1/12$.

Intuitively, suppose there are 1000 people in the population. Then 10 people have the disease, and 990 people do not have the disease. Among the 10 people who have the disease, 9 of them will be tested

positive. Among the 990 people that do not have the disease, 99 people will be tested positive. So in total, 99+9 = 108 people are tested positive. Among these 108 positive people, only 9 of them have the disease, so $P(disease|+) = 1/12$.

5.12 Mixture model and classification example

X is discrete and Y is continuous. Let $X = 1$ denote the gender is male. Let $P(X = 1) = p_X(1) = \rho, p_{Y|X}(y|1) = f_1(y), p_{Y|X}(y|0) = f_0(y)$. Then the marginal distribution is a mixture distribution:

$$\begin{aligned} p_Y(y) &= \sum_x p(x,y) \\ &= \sum_x p_X(x) p_{Y|X}(y|x) \\ &= p_X(1) p_{Y|X}(y|1) + p_X(0) p_{Y|X}(y|0) \\ &= \rho f_1(y) + (1-\rho) f_0(y). \end{aligned}$$

The posterior probability can be used for classification of Y:

$$\begin{aligned} P(X = 1|Y = y) &= p_{X|Y}(1|y) \\ &= \frac{p_X(1) p_{Y|X}(y|1)}{p_Y(y)} \\ &= \frac{\rho f_1(y)}{\rho f_1(y) + (1-\rho) f_0(y)}. \end{aligned}$$

Intuitively, suppose there are 1 million people in the population. Then the number of males is ρ million. Among these ρ million people, the proportion of people in $(y, y+\Delta y)$ is $f_1(y)\Delta y$. Thus there are $\rho f_1(y)\Delta y$ million males in $(y, y+\Delta y)$. Similarly, there are $(1-\rho) f_0(y)\Delta y$ females in $(y, y+\Delta y)$. Thus there are $(\rho f_1(y) + (1-\rho) f_0(y))\Delta y$ million people in $(y, y+\Delta y)$. The density is $\rho f_1(y) + (1-\rho) f_0(y)$.

Among the $(\rho f_1(y) + (1-\rho) f_0(y))\Delta y$ million people in $(y, y+\Delta y)$, the number of males is $\rho f_1(y)\Delta y$ million. Thus $P(male|y) = \rho f_1(y)/(\rho f_1(y) + (1-\rho) f_0(y))$.

5.13 Acceptance-rejection sampling example

X is continuous and Y is discrete. Steps: (a). Generate $X \sim g(x)$. (2) Return X if accepted, i.e. $Y = 1$, with probability $f(X)/(cg(X))$; or go back to (b) if rejected, i.e., $Y = 0$.

$$\begin{aligned} p_{X|Y}(x|1) &= \frac{p(x,1)}{p_Y(1)} \\ &= \frac{g(x) p_{Y|X}(1|x)}{\int_x g(x) p_{Y|X}(1|x)} \\ &= \frac{g(x)\frac{f(x)}{cg(x)}}{\int_x g(x)\frac{f(x)}{cg(x)}} = f(x). \end{aligned}$$

This is also related to missing data problem in statistics, as well as biased sampling.

The intuition is as follows. Under repetitions, step (a) generates a large sample of points $X_1, ..., X_n$ whose density is $f(x)$, i.e., the number of points in $(x, x+\Delta x)$ is $ng(x)\Delta x$. For these points to go through step (b), only a proportion $p(x) = f(x)/(cg(x))$ of points are accepted, so the number of points the survive step (b) is $ng(x)\Delta x p(x) = nf(x)\Delta x/c$. The total number of points left is $\sum_x nf(x)\Delta x/c = n/c$, so $1/c$ is the overall

acceptance rate. Among all these n/c points, the proportion of those in $(x, x+\Delta x)$ is $(nf(x)\Delta x/c)/(n/c) = f(x)\Delta x$. Thus the points left follow the desired density.

We may also let Ω be the region under the curve $cg(x)$. We can throw the points uniformly into Ω. Then the x coordinates of the points follow $g(x)$. Among all these point, we only accept those points under $f(x)$. We call the region under $f(x)$ to be A. On each vertical slice at x, the acceptance probability is $f(x)/(cg(x))$. The overall acceptance rate is $|A|/|\Omega| = 1/c$. Among those accepted points, the x coordinates of these points follow $f(x)$.

5.14 Bivariate normal example

Both X and Y are continuous. Consider the following model $X \sim \mathrm{N}(0,1)$ and $[Y|X=x] \sim \mathrm{N}(\rho x, 1-\rho^2)$. The joint density is

$$\begin{aligned} f(x,y) &= f_X(x) f_{Y|X}(y|x) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right) \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2+y^2-2\rho xy)\right]. \end{aligned}$$

The parameter ρ is actually the covariance between X and Y. The model is symmetric in (x,y). Thus $Y \sim \mathrm{N}(0,1)$ and $[X|Y=y] \sim \mathrm{N}(\rho y, 1-\rho^2)$.

Intuitively, $f(x,y)$ is the density of the scatterplot of $(X_i,Y_i) \sim f(x,y)$ independently for $i = 1,...,n$. $f(y|x)$ is the distribution of Y_i for those $X_i \in (x, x+\Delta x)$, which is a vertical slice. $f(x|y)$ is the distribution of X_i for those $Y_i \in (y, y+\Delta y)$, which is a horizontal slice. We may think of $f(y|x)$ as defining an equation or a mapping from X to Y, and $f(x|y)$ as defining the solution to the stochastic equation or the inverse of the stochastic mapping.

5.15 Shared cause property

Let Ω be the population of smokers. For smoker Ω, let $X(\omega) \in \{pipe, cigarette\}$ be the smoking habit of Ω, $Y(\omega) \in \{healthy, unhealthy\}$ be the health of Ω, and $Z(\omega) \in \{young, old\}$ be the age of Ω. Then for fixed x, y, z, the event $A = \{\omega : X(\omega) = x\}$ and the event $B = \{\omega : Y(\omega) = y\}$ are independent given the event $C = \{\omega : Z(\omega) = z\}$, i.e., within the same age group, the smoking habit of a random smoker has nothing to do with his or her health. C is the shared cause of A and B. However, A and B are not marginally independent, in fact, marginally or within the whole population of smokers, people who smoke cigarette seem to be healthier than those who smoke pipe. But this is caused by the fact that old people tend to smoke pipe while young people tend to smoke cigarette.

To be more specific, we can calculate

$$p(x,y) = \sum_z p(x,y,z) = \sum_z p(x,y|z)p(z) = \sum_z p(x|z)p(y|z)p(z).$$

using the rule of total probability. $p(x,y|z) = p(x|z)p(y|z)$ because of conditional independence. From $p(x,y)$, we can then calculate $p(x) = \sum_y p(x,y)$ again using the rule of total probability. Then we can calculate $p(y|x) = p(x,y)/p(x)$ using the definition of conditional probability.

To understand the above calculation, let us again count a population of 1 million smokers. Then $P(Z = young) \times 1\ million$ is the number of young smokers, and $P(Z = old) \times 1\ million$ is the number of old smokers. $P(X = cig, Y = healthy|Z = young)$ is the fraction of those who smoke cig and who are

healthy among young smokers. So $P(X = cig, Y = healthy|Z = young)P(Z = young)$ is the number (in million) of young healthy cig smokers. Similarly, $P(X = cig, Y = healthy|Z = old)P(Z = old)$ is the number (in million) of old healthy cig smokers. Thus $P(X = cig, Y = healthy|Z = young)P(Z = young) + P(X = cig, Y = healthy|Z = old)P(Z = old) = P(X = cig, Y = healthy)$ is the overall number (in million) of heathy cig smokers, pooled across different age groups, or $P(X = cig, Y = healthy)$ the proportion of heathy cig smokers within the whole population of 1 million smokers. Similarly you can calculate $P(X = x, Y = y)$ for other combinations of (x, y).

In order to calculate $P(X = cig, Y = healthy|Z = young)P(Z = young)$, we can decompose it using the chain rule, i.e., $P(Z = young)P(X = cig|Z = young)P(Y = healthy|X = cig, Z = young)$. The interpretation is as follows. The number of young smokers is $P(Z = young) \times 1\ million$. Among them, the fraction of cig smokers is $P(X = cig|Z = young)$. So if we multiply them, we get $P(Z = young)P(X = cig|Z = young) = P(X = cig, Z = young)$, which is the number (in million) of young cig smokers. $P(Y = healthy|X = cig, Z = young)$ is the fraction of healthy people among young cig smokers. So $P(Z = young)P(X = cig|Z = young)P(Y = healthy|X = cig, Z = young)$ is the number (in million) of young healthy cig smokers.

In the above example, the conditional independence assumption enables us to conclude that $P(Y = healthy|X = cig, Z = young) = P(Y = healthy|Z = young) = P(Y = healthy|X = pip, Z = young)$. That is, the fraction of healthy people among young cig smokers is the same as the fraction of healthy people among young pip smokers. In other words, smoking cig or pip has no effect whatsoever on whether one is healthy or not within the same age group. However, if we pool over the age groups, then smoking cig appears to be related to one being healthy, because young people tend to smoke cig.

5.16 Bayes net, directed graphical model

The repeated application of Markov property and the shared cause property enables us to build a Bayes net or graphical model or an expert system that involve many random variables. The Bayes net can serve as the basis for learning and inference.

A well-known toy example is the Asia example.

The above diagram shows a directed graphical model, for which we can write the joint distribution

$$p(v, s, t, l, b, c, x, d) = p(v)p(s)p(t|v)p(l|s)p(b|s)p(c|t, l)p(x|c)p(d|c, b).$$

For a query, e.g., $p(t|x, d, v) = p(t, x, d, v)/p(x, d, v)$, where both $p(t, x, d, v)$ and $p(x, d, v)$ can be calculated from the joint distribution by marginalizing out the irrelevant variables.

5.17 Causality

For a random smoker, let $X \in \{1 = cigarette, 0 = pipe\}$ be his or her smoking habit, let Y be his or her health. Suppose we observe that people who smoke cigarette are in general healthier than people who smoke pipe, can we draw the conclusion that smoking cigarette is better for health than smoking pipe? If someone cannot quit smoking, can we recommend him or her to smoke cigarette instead of pipe?

More formally, can we define $\mathbb{E}(Y|X = 1)$ as the causal effect of smoking cigarette, and $\mathbb{E}(Y|X = 0)$ as the causal effect of smoking pipe, so that $\mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$ is the causal effect of smoking cigarette versus smoking pipe?

Confounding variables. The problem with such a definition is that there may be confounding variables, such as $Z = age$. What if younger people tend to smoke cigarette and older people tend to smoke pipe? Then the difference between $\mathbb{E}(Y|X = 1)$ and $\mathbb{E}(Y|X = 0)$ may be due to the age difference between cigarette smokers and pipe smokers, instead of being caused by the smoking habit. So we need to define the causal effect more carefully.

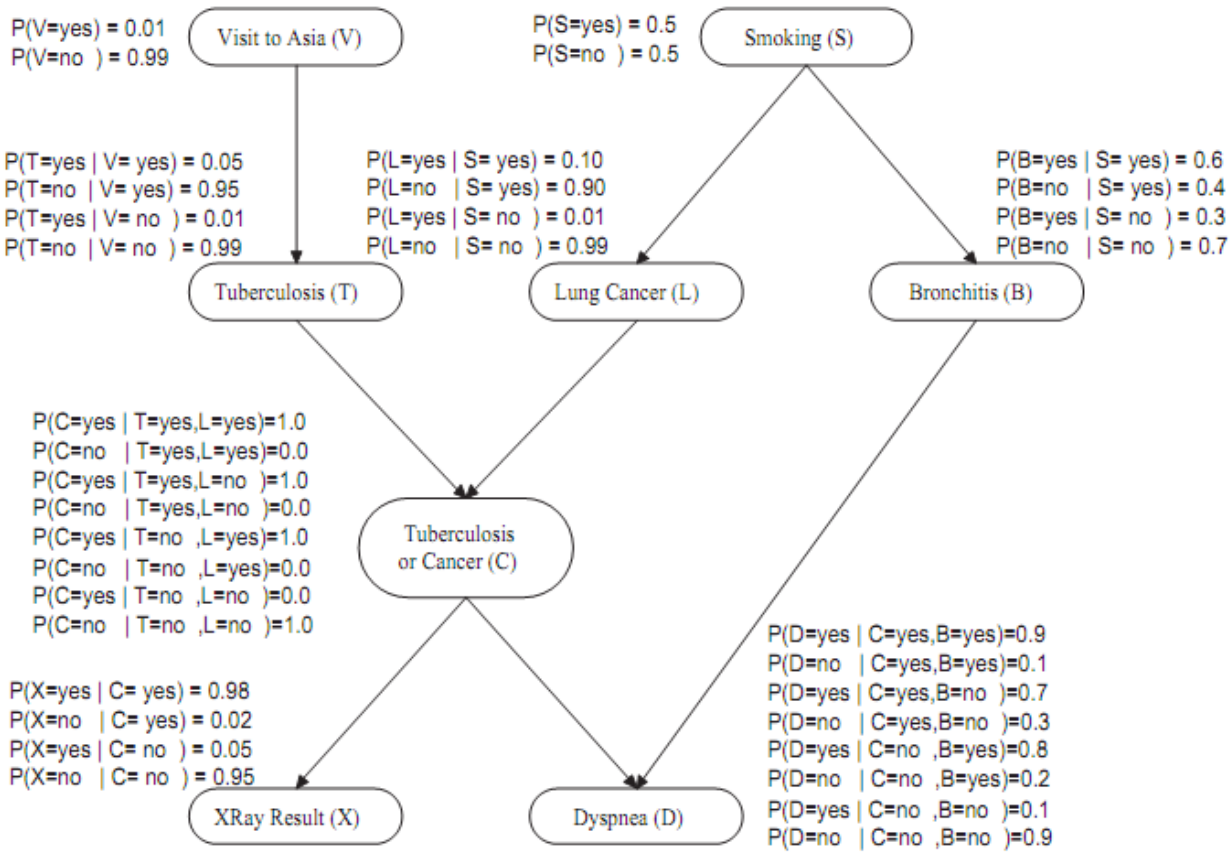

Figure 11: Asia example of Bayes net. Source: web.

Rubin's potential outcomes. Both Fisher and Neyman made important contributions to the issue of causality. But Rubin is the first one to formalize the conceptual framework of causal inference based on counterfactual potential outcomes. You may think causality must be a very profound topic. The truth is the opposite. Everyone knows how to do causal inference (except perhaps statisticians). Suppose Joe is having a bad cough, and he regrets that he has smoked so heavily recently. He says: if I had not smoked, I would not have this cough. See, Joe is doing a very valid causal inference with this counterfactual statement. We can do at least as well as Joe.

Pearl's structural equation and causal diagram. We can also treat this problem using Pearl's structural equations. (1) $Z = h_Z(\varepsilon_Z)$; (2) $X = h_X(Z, \varepsilon_X)$; and (3) $Y = h_Y(X, Z, \varepsilon_Y)$. These equations can be represented by a causal diagram, where there is an arrow from Z to X, and there are arrows from both Z and X to Y. We can map these structural equations to (1) $Z(s) = h_Z(\varepsilon_Z(s))$; (2) $X(s) = h_X(Z(s), \varepsilon_X(s))$; and (3) $Y(s) = h_Y(X(s), Z(s), \varepsilon_Y(s))$. The causal effect cannot be defined as $\mathbb{E}(Y|X = 1)$, but should be defined as $\mathbb{E}(Y|X \leftarrow 1)$ or $\mathbb{E}(Y|X = do(1))$, which corresponds to cutting the arrow from Z to X, so that (2) should be changed to $X \leftarrow 1$. The conditional expectation for this altered code $\mathbb{E}(Y|X \leftarrow 1)$ is the causal effect. In the altered code, (2) actually means $X(s) \leftarrow 1$, and (3) becomes $Y(s) = h_Y(X(s) = 1, Z(s), \varepsilon_Y(s))$, which is actually the potential outcome $Y_1(s)$ that we defined before in the Rubin's framework. So $\mathbb{E}(Y|X \leftarrow 1)$ is actually $\mathbb{E}(Y_1)$.

So the game is: if we repeat the original code 1 million times, and get 1 million copies of (X, Y, Z), whether we can predict what the distribution or average of Y if we repeat the altered code 1 million times? For the output from the original code, suppose we only keep those copies of Y with $X = 1$, and then computer

the average of these Y. Then this won't be the average of Y from the altered code. Even though both code share the same third line $Y = h_Y(X,Z,\varepsilon_Y)$, and we always plug in $X = 1$ in both code, the Z values from the two codes are very different. In the original code, the Z values that we keep correspond to $p(z|X=1)$. But in the altered code, the Z values correspond to $p(z)$ from $Z = h_Z(\varepsilon_Z)$.

For the original code, the copies of Y with $X = 1$ correspond to $\{Y_1(s) : X(s) = 1\}$. The values of Y produced by the altered code correspond to $\{Y_1(s), s \in S\}$. As we reasoned above, the two have different distributions unless the assignment X is independent of any possible confounding variables.

More formally, let $p(Y = y|Z = z)$ be the conditional distribution of Y given Z according to the original code. Let $\tilde{p}(Y = y|Z = z)$ be the distribution of Y according to the altered code. Then

$$\begin{aligned} p(Y=y|Z=z) &= \int p(Y=y,X=x|Z=z)dx \\ &= \int p(Y=y|X=x,Z=z)p(X=x|Z=z)dx. \end{aligned}$$

$$\begin{aligned} \tilde{p}(Y=y|Z=z) &= \int \tilde{p}(Y=y,X=x|Z=z)dx \\ &= \int \tilde{p}(Y=y|X=x,Z=z)\tilde{p}(X=x|Z=z)dx \\ &= \int p(Y=y|X=x,Z=z)p(X=x)dx, \end{aligned}$$

where $\tilde{p}(X = x|Z = z) = p(X = x)$ because Z is independent of X in the second code.

p and $\tilde{p}$ are the same if $p(X = x|Z = z) = \tilde{p}(X = x|Z = z) = p(X = x)$, i.e., when X and Z are independent.

6 Law of large numbers

6.1 Sample average converges to expectation

Let $X_1, X_2, ..., X_n \sim p(x)$ independently. Let $\mu = \mathbb{E}(X_i)$ and $\sigma^2 = \text{Var}(X_i)$. Let $\bar{X} = \sum_{i=1}^n X_i/n$. Then

$$\mathbb{E}(\bar{X}) = \frac{\mathbb{E}(\sum_{i=1}^n X_i)}{n} = \frac{\sum_{i=1}^n \mathbb{E}(X_i)}{n} = \frac{n\mu}{n} = \mu.$$

$$\text{Var}(\bar{X}) = \frac{\text{Var}(\sum_{i=1}^n X_i)}{n^2} = \frac{\sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

As $n \to \infty$, $\text{Var}(\bar{X}) \to 0$, thus $\bar{X} \to \mu$.

A more precise statement is, for any $\varepsilon > 0$, no matter how small it is, $P(|\bar{X} - \mu| < \varepsilon) \to 1$. This is a consequence of $\text{Var}(\bar{X}) \to 0$. This is called the weak law of large numbers. There is a stronger version which says that $P(\bar{X} \to \mu) = 1$.

One may wonder, didn't we just assume that expectation is the long run average? Actually we did define expectation as population average, but we did not define expectation as long run average. We only interpret the expectation as the limit of the long run average. The average is itself a random variable with diminishing variance. Consider a population of M people, and each person carries a number. The population average of these M numbers is μ. The population variance of these M numbers is σ^2. If we sequentially sample n people from the same population (with replacement). Then the number of possible sequences is M^n. For each sequence, we can calculate an average. So there are M^n averages or M^n numbers. The average of these M^n numbers is μ, but the variance of these M^n numbers is σ^2/n, and almost all of these M^n numbers are very close to μ.

6.2 Markov, Chebyshev, and weak law

Markov inequality. For a random variable $Z > 0$, for any $t > 0$, the tail probability $P(Z > t) \leq \mathbb{E}(Z)/t$. The proof is very simple. $P(Z > t) = \int_t^\infty f(z)dz$. $\mathbb{E}(Z) = \int zf(z)dz \geq \int_t^\infty zf(z)dz \geq \int_t^\infty tf(z)dz = tP(z > t)$.

Chebyshev inequality. $P(|X - \mu_X| > \varepsilon) \leq \mathbb{E}[(X - \mu_X)^2]/\varepsilon^2 = \mathrm{Var}(X)/\varepsilon^2$.

Weak law of large numbers. $P(|\bar{X} - \mu_X| > \varepsilon) \leq \mathrm{Var}(\bar{X})/\varepsilon^2 = \sigma_X^2/(n\varepsilon^2) \to 0$.

6.3 Strong law of large number

Let the tail event

$$A_n = \{(x_1, x_2, ...) : |\bar{x}_n - \mu| > \varepsilon\}.$$

For uniform distribution, A_n is the off-diagonal part of the n-dimensional cube $[0,1]^n$. The weak law of large number claims that $P(A_n) \to 0$ as $n \to \infty$, i.e., the volume of the diagonal piece, no matter how thin it is, i.e., no matter how small ε is, goes to 1. In the weak law, when we measure the volume, we always fix n, which is finite. After we measure the volume, we let $n \to \infty$.

The strong law goes a step further. We consider

$$B = \{(x_1, x_2, ...) : \bar{x}_n \to \mu\}$$

where the sample space consists of all the infinite sequences. In the case of uniform distribution, the sample space is the infinite dimensional cube, $[0,1]^\infty$, and B is the diagonal piece of this infinite dimensional cube. The strong law of large number claims that $P(B) = 1$, i.e., the volume of the diagonal piece is 1. In other words, almost all the points in the infinite dimension cube (the coordinates of each point is an infinite sequence $(x_1, x_2, ...)$) belong to the diagonal piece. The points that do not belong to the diagonal piece, i.e., the sequences $(x_1, x_2, ...)$ whose $\bar{x}_n$ do not converge, are so few that they are like the rational points in $[0,1]$, whose total length is 0.

The meaning of $\bar{x}_n \to \mu$ is as follows. For any $\varepsilon > 0$, there exists an $N \geq 1$, such that for any $n \geq N$, $|\bar{x}_n - \mu| \leq \varepsilon$. The opposite of this statement is, for any $\varepsilon > 0$, for any $N \geq 1$, there exists $n \geq N$, such that $|\bar{x}_n - \mu| > \varepsilon$. Let $A = B^c$, then

$$A = \cap_\varepsilon \cap_{N=1}^\infty \cup_{n=N}^\infty A_{n,\varepsilon},$$

where ε may take values in a discrete list that goes to 0, such as $\{1/k, k = 1, 2, ...\}$. For uniform distribution, A defines the off-diagonal piece of the infinite dimensional cube. The strong law of large number says that $P(A) = 0$.

The strong law is stronger than the weak law, which says $P(A_n) \to 0$, where $P(A_n)$ is measured on the n-dimensional cube, and the convergence takes place outside P. In strong law, the convergence takes place inside P, and the probability is measured on the infinite dimensional cube.

Measure theory enables us to measure A in the infinite dimensional cube, via the infinite additivity of the measure. A belongs to the σ-algebra generated by the basic sets or basic statements such as A_n, which are subsets of finite dimensional cubes.

6.4 Borel-Cantelli Lemma

The door to the strong law of large number is the Borel-Cantelli Lemma, which says that if

$$\sum_{n=1}^\infty P(A_n) < \infty,$$

then

$$P(\cap_{N=1}^\infty \cup_{n=N}^\infty A_n) = 0,$$

i.e., the probability that A_n happens infinitely often (i.o.) is 0. This is because $P(\cap_{N=1}^{\infty}\cup_{n=N}^{\infty}A_n) \le P(\cap_{n=N}^{\infty}A_n) \le \sum_N^{\infty}P(A_n) \to 0$ as $N \to \infty$.

The condition $\sum_{n=1}^{\infty}P(A_n) < \infty$ is sufficient for the fact that $P(A_n\ i.o.) = 0$. It is also necessary if A_n are independent.

If for any ε, $P(\cap_{N=1}^{\infty}\cup_{n=N}^{\infty}A_{n,\varepsilon}) = P(A_{n,\varepsilon}\ i.o.) = 0$, then $P(A) = P(\cap_{\varepsilon}\cap_{N=1}^{\infty}\cup_{n=N}^{\infty}A_{n,\varepsilon}) = 0$, so we have the strong law of large number.

For $X_1, X_2, ... \sim$ Bernoulli(p) independently, we have $P(A_n) \le 2\exp(-2n\varepsilon^2)$, so $\sum_{i=1}^{n}P(A_n) < \infty$, thus according to Borel-Cantelli, the strong law is true, and $P(\bar{X}_n \to p) = 1$. For bounded random variables, according to Hoeffding, we also have the strong law.

Technically, the strong law is hinged on $\sum_{n=N}^{\infty}P(A_n) \to 0$ as $N \to \infty$, while the weak law is about $P(A_n) \to 0$.

6.5 ℓ_2 strong law

We may assume $X \ge 0$. Otherwise, we can always write $X = X_+ - X_-$, where $X_+ = \max(0, X)$ and $X_- = \max(0, -X)$, and deal with X_+ and X_- separately.

Assuming $\sigma^2 = \mathrm{Var}(X) < \infty$, according to Chebyshev, we have $P(A_{n,\varepsilon}) \le \sigma^2/(n\varepsilon^2)$. However, $\sum_{n=1}^{\infty}1/n = \infty$, thus we cannot apply Borel-Cantelli directly to prove the ℓ_2 strong law. However, this does not mean that the strong law does not hold, because $\sum_{n=1}^{n}P(A_n) < \infty$ is sufficient but not necessary. It is necessary only if A_n are independent. In our case, this is clearly not the case.

We know that $\sum_{n=1}^{\infty}1/n^2 < \infty$, so we may think of A_{n^2} first. According to Borel-Cantelli, $P(A_{n^2}\ i.o.) = 0$, and we have strong law of large number for $\bar{X}_{n^2}$, i.e., $P(\bar{X}_{n^2} \to \mu) = 1$. For any given sequence $(x_1, x_2, ...)$, for any $n^2 < m < (n+1)^2$,

$$\bar{x}_{n^2}\frac{n^2}{(n+1)^2} \le \frac{\sum_{i=1}^{n^2}x_i}{(n+1)^2} \le \frac{\sum_{i=1}^{n^2}x_i}{m} \le \frac{\sum_{i=1}^{m}x_i}{m} \le \frac{\sum_{i=1}^{(n+1)^2}x_i}{m} \le \frac{\sum_{i=1}^{(n+1)^2}x_i}{n^2} = \bar{x}_{(n+1)^2}\frac{(n+1)^2}{n^2}.$$

If $\bar{x}_{n^2} \to \mu$, then we must have $\bar{x}_m \to \mu$ as $m \to \infty$. Thus the event $\{(x_1, x_2, ...) : \bar{x}_{n^2} \to \mu\}$ is the same as the event $\{(x_1, x_2, ...) : \bar{x}_n \to \mu\}$. So $P(\bar{X}_n \to \mu) = 1$. We say $\bar{X}_n \to \mu$ almost surely (a.s.) or almost everywhere (a.e.).

The strong law also holds for ℓ_1 case, i.e., we only need to assume $\mathbb{E}(|X|) < \infty$. In order to prove this result, we need even sparser sequence $\bar{X}_{\gamma^n}$ for $\gamma > 1$, and we need to deal with truncated random variable $X_M = 1(X \le M)X$.

7 Large deviation

7.1 Chernoff trick and large deviation upper bound

The original Markov bound is too loose. However, we can sharpen it by Chernoff's trick:

$$\begin{aligned} P(X > t) &= P(e^{\lambda X} > e^{\lambda t}) \\ &\le \frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda t}} = \frac{M_X(\lambda)}{e^{\lambda t}}, \end{aligned}$$

where $\lambda > 0$.

So to bound the tail probability for $\bar{X}$, we have

$$\begin{aligned} P(\bar{X} > t) &= P(\sum_{i=1}^{n} X_i > nt) \leq \frac{\mathbb{E}(e^{\lambda \sum_{i=1}^{n} X_i})}{e^{\lambda nt}} \\ &= \prod_{i=1}^{n} \mathbb{E}(e^{\lambda X_i})/e^{n\lambda t} = \frac{M_X(\lambda)^n}{e^{\lambda t}} \\ &= \exp[-n(\lambda t - \log M_X(\lambda))] \\ &\leq e^{-nI_X(t)}, \end{aligned}$$

where $I_X(t) = \sup_{\lambda \geq 0}[\lambda t - \log M_X(\lambda)]$. To be more specific, let $\hat{\lambda}$ be the λ where the supreme is achieved, then $I_X(t) = \hat{\lambda} t - \log M_X(\hat{\lambda})$. Since $I_X(0) = 0$, $I_X(t) \geq 0$. Here we have to assume that $M_X(\lambda)$ exists for an interval $\lambda \in [0, c)$ for some constant $c > 0$.

7.2 Moment generating function

$M_X(\lambda) = \mathbb{E}(e^{\lambda X})$ is the moment generating function. It has the following properties:

(1) $M_X(0) = 1$, $M_X'(0) = \mathbb{E}(X)$. $M_X''(0) = \mathbb{E}(X^2)$. $M_X^{(k)}(0) = \mathbb{E}(X^k)$.

(2) For iid $X_1, ..., X_n \sim f_X(x)$, $M_{\bar{X}}(\lambda) = \mathbb{E}(e^{\lambda \bar{X}}) = \mathbb{E}(e^{\lambda (X_1 + ... + X_n)/n}) = \prod_{i=1}^{n} \mathbb{E}(e^{\lambda X_i/n}) = M_X(\lambda/n)^n$.

Later we shall also use the notation $Z(\lambda) = M_X(\lambda)$. The first and second derivatives of $\log Z(\lambda)$ are expectation and variance of a distribution.

7.3 Importance sampling, exponential tilting, and lower bound

In order to calculate the tail probability, such as $P(X > 10)$ for $X \sim \mathrm{N}(0, 1)$ using Monte Carlo method, we can repeatedly generate X from $\mathrm{N}(0, 1)$ and count the frequency of $X > 10$. But for a limited number of repetitions, we may never encounter any occurrence of $X > 10$. One method to remedy this situation is to generate $X \sim \mathrm{N}(10, 1)$, then half of the times we will have $X > 10$. Then we can obtain

$$\begin{aligned} P(X > 10) &= \int_{x>10} p(x)dx = \int_{x>10} \frac{p(x)}{\hat{p}(x)} \hat{p}(x)dx \\ &= \mathbb{E}_{\hat{p}}\left[1(X > 10)\frac{p(X)}{\hat{p}(X)}\right] \\ &= \mathbb{E}_{\hat{p}}[1(X > 10)w(X)] \end{aligned}$$

where $p(x)$ is $\mathrm{N}(0, 1)$, $\hat{p}(x)$ is $\mathrm{N}(10, 1)$, $w(x) = p(x)/\hat{p}(x)$ is the importance weight to account for the fact that X is sampled from $\hat{p}$ instead of the original p. In Monte Carlo, the above expectation can be approximated by long-run average by sampling from $\hat{p}(x)$.

We can apply the above idea to calculating $P(\bar{X} > t)$. Let

$$p_\lambda(x) = \frac{1}{Z(\lambda)} \exp(\lambda x) p(x),$$

where

$$Z(\lambda) = \int \exp(\lambda x) p(x) = \mathbb{E}_p[\exp(\lambda X)] = M_X(\lambda)$$

is the normalizing constant. $p_\lambda(x)$ is an exponential tilting of $p(x)$. We want to choose λ so that $\mathbb{E}_\lambda(X) = t$.

Let the resulting λ be $\hat{\lambda}$. Then

$$\begin{aligned} P(\bar{X} > t) &= \mathbb{E}_{\hat{\lambda}} \left[1(\bar{X} > t) \frac{\prod_{i=1}^n p(X_1, ..., X_n)}{\prod_{i=1}^n \hat{p}(X_1, ..., X_n)} \right] \\ &= \mathbb{E}_{\hat{\lambda}} \left[1(\bar{X} > t) \exp[-(\hat{\lambda} \sum_{i=1}^n X_i - n \log Z(\hat{\lambda}))] \right] \end{aligned}$$

Since $\mathbb{E}_{\hat{p}}(X) = t$, if we sample $X_1, ..., X_n \sim \hat{p}(x)$ independently, then $\sum_{i=1}^n X_i/n \to t$, so the above quantity can be lower bounded by a constant times $\exp[-n(\hat{\lambda} t - \log Z(\hat{\lambda})] = \exp[-nI_X(t)]$. Thus the rate of the large deviation upper bound $I_X(t)$ is also the rate of the large deviation lower bound, so $I_X(t)$ is the exact rate.

7.4 Sub-gaussian distribution

For the normal distribution $X \sim \mathrm{N}(0,1)$, the moment generating function

$$M(\lambda) = \mathbb{E}[\exp(\lambda X)] = e^{\lambda^2/2}.$$

We define a random variable X to be sub-Gaussian if its moment generating function

$$M(\lambda) = \mathbb{E}[\exp(\lambda X)] \leq e^{\lambda^2/2}.$$

The Chernoff bound is

$$\begin{aligned} P(X > t) &= P(e^{\lambda X} > e^{\lambda t}) \\ &\leq \frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda t}} = \frac{M(\lambda)}{e^{\lambda t}} \\ &\leq \exp[\lambda^2/2 - \lambda t]. \end{aligned}$$

We can minimize the above upper bound over λ. The minimum is achieved at $\lambda = t$, and the bound is

$$P(X > t) \leq e^{-t^2/2},$$

or

$$P(|X| > t) \leq 2e^{-t^2/2}.$$

7.5 Gibbs distribution, partition function and derivatives

The distribution $p_\lambda(x)$ is called the Gibbs distribution in statistical mechanics. It is called exponential family model in statistics, and energy-based model in machine learning. The large deviation rate is the deepest place to find such a distribution.

The normalization constant $Z(\lambda)$ is called the partition function of the Gibbs distribution. Two important properties of $Z(\lambda)$ is that

$$\frac{d}{d\lambda} \log Z(\lambda) = \frac{Z'(\lambda)}{Z(\lambda)} = \frac{1}{Z(\lambda)} \int e^{\lambda x} x p(x) dx = \int x p_\lambda(x) dx = \mathbb{E}_\lambda(X),$$

and

$$\frac{d^2}{d\lambda^2} \log Z(\lambda) = \frac{Z''(\lambda)}{Z(\lambda)} - \frac{Z'(\lambda)^2}{Z(\lambda)^2} = \mathbb{E}_\lambda(X^2) - \mathbb{E}_\lambda(X)^2 = \mathrm{Var}_\lambda(X).$$

Thus in the large deviation upper bound where we want to maximize $L(\lambda) = \lambda t - \log Z(\lambda)$, we can set $L'(\lambda) = 0$, which is $\mathbb{E}_\lambda(X) = t$, so the supreme is achieved by $\hat{\lambda}$ in the lower bound. $L(\lambda)$ is a concave function because $L'(\lambda) = -\mathrm{Var}_\lambda(X) \leq 0$, so $\hat{\lambda}$ is the maximum.

7.6 Hoeffding inequality, concentration of measure

The large deviation upper bound $P(\bar{X} > t) \le \exp[-n(\lambda t - \log Z(\lambda))]$ or more specifically $\log Z(\lambda)$, depends on the original distribution p. We can find a more general bound that only depends on some generic properties of p. For simplicity, let us assume $\mathbb{E}_p(X) = 0$. If we further assume that X is bounded, i.e., $X \in [a,b]$, then we have the Hoeffding inequality.

Recall that $Z(\lambda)$ is the partition function of the tilted distribution $p_\lambda(x)$ discussed above. We take the second order Taylor expansion of $\log Z(\lambda)$ at $\lambda = 0$ ($p_0(x) = p(x)$),

$$\begin{aligned} \log Z(\lambda) &= \log Z(0) + \frac{d}{d\lambda}\log Z(\lambda)|_{\lambda=0}\lambda + \frac{1}{2}\frac{d^2}{d\lambda^2}\log Z(\lambda)|_{\lambda=\lambda_*}\lambda^2 \\ &= \frac{1}{2}\mathrm{Var}_{\lambda_*}(X)\lambda^2 \le \frac{1}{8}(b-a)^2\lambda^2, \end{aligned}$$

because $Z(0) = 1$, $d\log Z(\lambda)/d\lambda = \mathbb{E}_\lambda(X) = 0$ for $\lambda = 0$, since $\mathbb{E}_p(X) = 0$, and

$$\begin{aligned} d^2\log Z(\lambda)/d\lambda^2 &= \mathrm{Var}_\lambda(X) \\ &= \mathrm{Var}_\lambda[X - (b+a)/2] \\ &\le \mathbb{E}_\lambda[(X-(b+a)/2)^2] \le [(b-a)/2]^2. \end{aligned}$$

Thus $\lambda t - \log Z(\lambda) \ge \lambda t - \lambda^2(b-a)^2/8$, which reaches its maximum at $\lambda = 4t/(b-a)^2$, and the maximum is $2t^2/(b-a)^2$. So

$$P(\bar{X} > t) \le \exp\left[-\frac{2nt^2}{(b-a)^2}\right].$$

For $X_1, ..., X_n \sim \text{Bernoulli}(p)$, applying the Hoeffding's equation to $X_i - p$, which is bounded by $[a = -p, b = 1-p]$, we have

$$P(|\bar{X} - p| > \varepsilon) \le 2\exp(-2n\varepsilon^2).$$

There are other similar bounds, such as Bernstein inequality, which also depends on $\sigma^2 = \mathrm{Var}_p(X)$. These inequalities are called concentration inequalities. They underly the phenomenon called concentration of measure, i.e., for iid $X_1, ..., X_n$, the probability distribution of $X_1, ..., X_n$ concentrates on $\{x : |\frac{1}{n}\sum_{i=1}^n x_i - \mu| < \varepsilon\}$. For uniform distribution, it means the diagonal piece occupies almost the whole volume.

8 Central limit theorem

8.1 Small deviation

Suppose $X_1, ..., X_n \sim p(x)$ independently. Let $\mu = \mathbb{E}(X_i)$ and $\sigma^2 = \mathrm{Var}(X_i)$. We have shown that for $\bar{X} = \sum_{i=1}^n X_i/n$, $\mathbb{E}(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \sigma^2/n$. The law of large numbers claims that $\bar{X} \to \mu$. The central limit theorem is a refinement of the law of large numbers. It says that $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n}) \to N(0,1)$ no matter what the distribution $p(x)$ is. Or in other words, $\sqrt{n}(\bar{X} - \mu) \to N(0, \sigma^2)$. According to the law of large numbers, $\bar{X} - \mu \to 0$. We magnify this difference by $\sqrt{n}$. Then the magnified difference converges to a normal distribution.

The normal approximation of binomial distribution is a special case of the central limit theorem. A binomial random variable is the summation of Bernoulli random variables. So if $X_i \sim \text{Bernoulli}(1/2)$, then $X = \sum_{i=1}^n X_i \sim \text{Binomial}(n, 1/2)$. Then $Z = (X - n/2)/(\sqrt{n}/2) \sim \text{N}(0,1)$ in the limit.

Large deviation is about $P(|\bar{X}_n - \mu| > \varepsilon)$. This is large deviation no matter how small ε is, because $\mathrm{Var}(\bar{X}_n - \mu) = \sigma^2/n$. The central limit theorem enables us to calculate $P(|\bar{X}_n - \mu| > \varepsilon/\sqrt{n})$, which is about small deviation of order $1/\sqrt{n}$.

8.2 Moment generating function

Suppose $\mathbb{E}(X) = 0$, $\mathrm{Var}(X) = 1$ from now on. Then $\mathbb{E}(\bar{X}) = 0$, $\mathrm{Var}(\bar{X}) = 1/n$. Let $Z_n = \sqrt{n}\bar{X}$. Then $\mathbb{E}(Z_n) = 0$, $\mathrm{Var}(Z_n) = 1$.

$$\begin{aligned}
M_n(\lambda) &= \mathbb{E}(e^{\lambda Z_n}) \\
&= \mathbb{E}(e^{\lambda(X_1+\ldots+X_n)/\sqrt{n}}) \\
&= \prod_{i=1}^{n} \mathbb{E}(e^{\lambda X_i/\sqrt{n}}) \\
&= M_X(\lambda/\sqrt{n})^n \\
&= (1 + M_X'(0)\frac{\lambda}{\sqrt{n}} + \frac{1}{2}M_X''(0)\frac{\lambda^2}{n} + o(\frac{1}{n}))^n \\
&= (1 + \frac{1}{2}\frac{\lambda^2}{n} + o(\frac{1}{n}))^n \\
&\to e^{\lambda^2/2}.
\end{aligned}$$

For a normal random variable $Z \sim \mathrm{N}(0,1)$,

$$\begin{aligned}
M(\lambda) &= \mathbb{E}(e^{\lambda Z}) \\
&= \int e^{\lambda z}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \\
&= \int \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-\lambda)^2}e^{\lambda^2/2} \\
&= e^{\lambda^2/2}.
\end{aligned}$$

So $M_n(\lambda) \to M(\lambda)$. Thus $Z_n \to \mathrm{N}(0,1)$ in distribution, i.e., for any z, $P(Z_n \le z) \to P(Z \le z)$.

For the above proof to hold, we need (1) Existence. $M(\lambda)$ exists for λ within an interval that contains 0. (2) Uniqueness, the moment generating function uniquely determines the cumulative density function. (3) Continuity, if $M_{X_n}(\lambda) \to M_X(\lambda)$ for every λ within an interval that contains 0, then the cumulative density function of X_n converges to that of X at every point.

8.3 Characteristic function

For distributions with long tails, $M(\lambda)$ may not exist. So it is better to use characteristic function by letting $\lambda = i\omega$, i.e., for a random variable $X \sim f_X(x)$, define the characteristic function

$$\phi_X(\omega) = \mathbb{E}(e^{i\omega X}) = \int f_X(x)e^{i\omega x}dx.$$

The characteristic function is actually the Fourier transform of the density function. The three properties mentioned in the previous subsection can be established for characteristic functions. For standard normal random variable $Z \sim \mathrm{N}(0,1)$, $\phi_Z(t) = e^{(i\omega)^2/2} = e^{-\omega^2/2}$, and $\phi_n(\omega) \to e^{-\omega^2/2}$, which can be proved in the same manner as above. So $Z_n \to \mathrm{N}(0,1)$ in distribution.

8.4 Convolution, Fourier transform and smoothing

To understand the effect of averaging, consider a bimodal $f(x)$, so that for $X \sim f(x)$, half of the time X is large and half of the time X is small. If we have $X_1, X_2 \sim f(x)$ independently, then we have the following

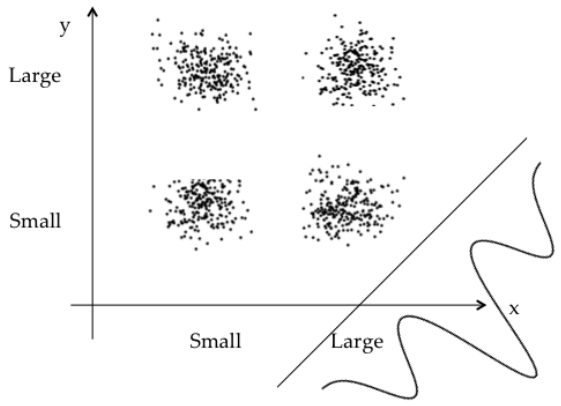

Figure 12: Averaging tends to reduce the variance and smooth the distribution, which leads to the law of large number and central limit theorem respectively.

four equally likely combinations (large, large), (large, small), (small, large), (small, small), so the average $\bar{X} = (X_1 + X_2)/2$ will be large, medium, medium, small respectively. Thus half of the time $\bar{X}$ will be medium, 1/4 of times $\bar{X}$ is small, and $1/4$ of times $\bar{X}$ is large. So the variability of $\bar{X}$ is reduced because there is more chance that X_1 and X_2 will balance each other, moreover, the distribution of $\bar{X}$ is smoother than the original distribution of X. The reduction of variance leads to the law of large number, and the smoothing of the distribution leads to the central limit theorem.

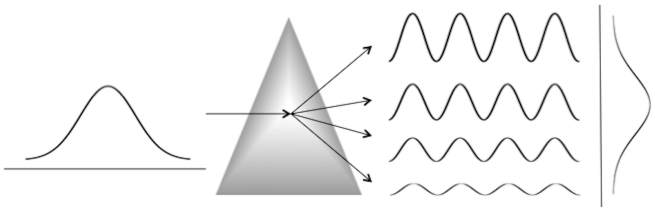

Figure 13: Representing a function $f(x)$ in the original domain by its Fourier components in the frequency domain, as if a composite color is decomposed into pure colors of different frequencies.

For two random variables X and Y that follow $f_X(x)$ and $f_Y(y)$ independently, their sum $Z = X + Y$ follows a density $f_Z(z)$ that can be obtained as follows.

$$\begin{aligned} f_Z(z) &= \frac{P(Z = X + Y \in (z, z+\Delta z))}{\Delta z} \\ &= \frac{\sum_x P(Y \in (y, y+\Delta y), Z \in (z, z+\Delta z))}{\Delta z} \\ &= \int f_X(z-y) f_Y(y) dy \\ &= \mathbb{E}_Y[f_X(z-Y)], \end{aligned}$$

so $f_Z(z)$ is obtained by smoothing $f_X(z)$ by local averaging according to f_Y. This is the reason that summation tends to smooth the distribution. This local averaging is a smoothing operation or a convolution operation.

The characteristic function $\phi_X(\omega)$ is the Fourier transform of $f_X(x)$. The Fourier transform decomposes $f_X(x)$ into different frequency components (like pure colors), and the amplitude of each frequency ω is given by $\phi_X(\omega)$. A key property of the Fourier transform can be seen very easily in the probability language

$$\phi_Z(\omega) = \mathbb{E}[e^{i\omega(X+Y)}] = \mathbb{E}[e^{i\omega X}]\mathbb{E}[e^{i\omega Y}] = \phi_X(\omega)\phi_Y(\omega).$$

In the original domain, $f_Z(z) = \int f_X(z-x)f_Y(y)dy = \int f_X(x)f_Y(z-x)dx$ is obtained by convolution, which is very complicated. In the frequency domain, $\phi_Z(\omega) = \phi_X(\omega)\phi_Y(\omega)$ is obtained by simple point-wise product. Since $|\phi_X(\omega)| \leq \mathbb{E}[|e^{i\omega X}|] = 1$, the convolution reduces the magnitudes of the frequency components and makes the resulting distribution smoother than the original distributions.

8.5 Lindeberg method

The proof based on the characteristic function or the Fourier transform is a global method, because each frequency component involves the whole distribution of $Z_n = \sum_{i=1}^n X_i/\sqrt{n}$. We would prefer a local method that deals with the effect of each X_i or $X_i/\sqrt{n}$. The Lindeberg method is such a local method.

Let $V_1, V_2, ...$ be iid $\mathrm{N}(0,1)$ random variables. Then $Z = \sum_{i=1}^n V_i/n$ is also a $\mathrm{N}(0,1)$ random variable. Consider a test function f that belongs to a certain class of functions $\mathscr{F}$, we want to prove $\mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)]$ for any $f \in \mathscr{F}$. For now, we assume that $\mathscr{F}$ consists of functions with finite third derivative, and that $\mathbb{E}[|X|^3] < \infty$.

The Lindeberg scheme sequentially flips the sequence $(X_1, ..., X_n)$ to $(V_1, ..., V_n)$. For instance, for $n = 3$, we can write

$$\begin{aligned} f(Z_n) - f(Z) &= f\left((X_1+X_2+X_3)/\sqrt{n}\right) - f\left((X_1+X_2+V_3)/\sqrt{n}\right) \\ &+ f\left((X_1+X_2+V_3)/\sqrt{n}\right) - f\left((X_1+V_2+V_3)/\sqrt{n}\right) \\ &+ f\left((X_1+V_2+V_3)/\sqrt{n}\right) - f\left((V_1+V_2+V_3)/\sqrt{n}\right). \end{aligned}$$

In general, let $S_i = X_1 + ... + X_{i-1} + V_{i+1} + ... + V_n$. Then each line is $f((S_i+X_i)/\sqrt{n}) - f((S_i+V_i)/\sqrt{n})$.

$$|\mathbb{E}[f(Z_n)] - \mathbb{E}[f(Z)]| \leq \sum_{i=1}^n |\mathbb{E}[f(S_i+X_i)/\sqrt{n})] - \mathbb{E}[f(S_i+V_i)/\sqrt{n}]|.$$

Let $a_i = S_i/\sqrt{n}$,

$$f(S_i+X_i)/\sqrt{n}) = f(a_i) + \frac{1}{\sqrt{n}}f'(a_i)X_i + \frac{1}{2n}f''(a_i)X_i^2 + R_1,$$
$$f(S_i+V_i)/\sqrt{n}) = f(a_i) + \frac{1}{\sqrt{n}}f'(a_i)V_i + \frac{1}{2n}f''(a_i)V_i^2 + R_2.$$

$\mathbb{E}(X_i) = \mathbb{E}(V_i) = 0$, $\mathrm{Var}(X_i) = \mathrm{Var}(V_i) = 1$. So the above two quantities equal to each other in expectation, except for the remainders R_1 and R_2. Let $|f^{(3)}| = \sup_x |f^{(3)}(x)|$, then $|R_1| \leq |f^{(3)}||X_i|^3/(6n^{3/2})$ and $|R_2| \leq |f^{(3)}||V_i|^3/(6n^{3/2})$, and

$$|\mathbb{E}[f(S_i+X_i)/\sqrt{n}] - \mathbb{E}[f(S_i+V_i)/\sqrt{n}]| \leq |R_1| + |R_2| \leq |f^{(3)}|(\mathbb{E}(|X|^3) + \mathbb{E}(|V|^3))/(6n^{3/2}).$$

Sum the above quantities for $i = 1, ..., n$, we have

$$|\mathbb{E}[f(Z_n)] - \mathbb{E}[f(Z)]| \leq |f^{(3)}|(\mathbb{E}(|X|^3) + \mathbb{E}(|V|^3))/(6\sqrt{n}) \to 0.$$

The above bound is called Berry-Essen bound for the central limit theorem. To connect to cumulative density functions of Z_n and Z, we need the testing function $f_z(x) = 1(x \leq z)$. We can construct a sequence $f_k \in \mathscr{F}$, and let $f_k \to f_z$.

8.6 Stein method

Another local method is the Stein method, which is based on the observation that for a random variable $Z \sim \mathrm{N}(0,1)$,

$$\mathbb{E}[f'(Z)] = \int f'(z)\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz = f(z)\frac{1}{\sqrt{2\pi}}e^{-z^2/2}|_{-\infty}^{\infty} + \int f(z)z\frac{1}{\sqrt{2\pi}}e^{-z^2/2} = \mathbb{E}[Zf(Z)],$$

assuming that the tail of $f(z)\frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ goes to 0 as $z \to \infty$ or $-\infty$. If the distribution of Z_n converges to the distribution of Z, then supposedly $\mathbb{E}[f'(Z_n) - Z_n f(Z_n)]$ should be small. We can then bound the latter, which only involves $Z_n = \sum_{i=1}^n X_i/\sqrt{n}$ but does not involve Z, so the analysis can be easy. More specifically, when we want to bound $\mathbb{E}[h(Z_n)] - \mathbb{E}[h(Z)]$ for some testing function h, we may find another function f so that

$$f'(x) - xf(x) = h(x) - \mathbb{E}[h(Z)].$$

Solving the above differential equation gives us such an f that is well behaved (which will not be the case if we replace $\mathbb{E}[h(Z)]$ by another quantity). Then we can bound $\mathbb{E}[f'(Z_n) - Z_n f(Z_n)]$. Even though f depends on h, the bound only requires some generic properties of f, instead of the specific properties of h. This method turns out to be very powerful in that it can deal with weakly dependent X_i, as well as more general notion of summation, i.e., replacing $\sum_i X_i$ by $S(X_1,...,X_n)$, where the influence of each X_i on S is limited. The concentration of measure can also be generalized to such settings.

9 Information theory

9.1 Equipartition property and entropy

For a discrete random variable $X \sim p(x)$, the entropy of p measures the randomness of p: $\text{entropy}(p) = \mathbb{E}[-\log X] = -\sum_x p(x)\log p(x)$.

Let $X_1,...,X_n \sim p(x)$ independently,

$$\begin{aligned}\frac{1}{n}\log_2 P(X_1,...,X_n) &= \frac{1}{n}\sum_{i=1}^n \log_2 p(X_i) \\ &\to \mathbb{E}[\log_2 p(X)] = -\text{entropy}(p).\end{aligned}$$

So in the limit, $P(X_1,...,X_n) \approx 2^{-n\text{entropy}(p)}$. The probability of a random sequence is a constant, and it is as if the sequence $X_1,...,X_n$ is drawn from a uniform distribution over a set.

In fact, we can define the typical set

$$\Omega_{n,\varepsilon} = \{(x_1,...,x_n) : |\frac{1}{n}\sum_{i=1}^n \log p(x_i) - \text{entropy}(p)| \le \varepsilon\},$$

then the distribution of $X_1,...,X_n$ is essentially a uniform distribution on $\Omega_{n,\varepsilon}$, and $|\Omega_{n,\varepsilon}|$ is about $2^{n\text{entropy}(p)}$. In other words, the distribution of $X_1,...,X_n$ concentrates uniformly on $\Omega_{n,\varepsilon}$. This property is called the equipartition property.

9.2 Coding and entropy

Consider a distribution $p(x)$ on an alphabet A,B,C,D, $p(A) = 1/2$, $p(B) = 1/4$, $p(C) = 1/8$, $p(D) = 1/8$. We can generate $X \sim p(x)$ using coin flipping. We flip a coin, if it is head, we output A. If it is tail, we continue to flip. If it is head, we output B. If it is tail, we continue to flip. If it is head, we output C. If it

x	A	B	C	D
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
$-2log_2 p(x)$	1	2	3	3
$code$	1	01	001	000

Figure 14: Distribution, coin flipping, and code

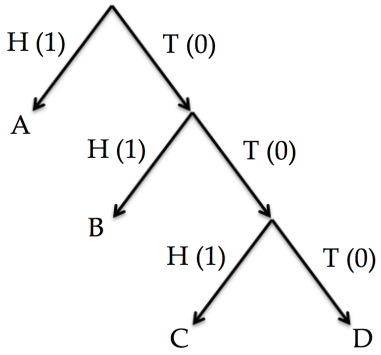

Figure 15: Coin flipping and coding

is tail, we output D. So $A = H$, $B = TH$, $C = TTH$, $D = TTT$. The expected number of coin flippings is entropy(p).

We can also change the above coin flippings into a binary code $A = 1$, $B = 01$, $C = 001$, $D = 000$. For a sequence generate from $p(x)$, we can code it and transmit it as a binary sequence. There are two important observations. (1) We can recover the original sequence from the binary sequence. This is because the binary code of any letter is not the beginning part of the binary code of any other letter. That is, the code is prefix code. This is because in the coin flipping experiment, when we output a letter, the experiment then stops. So the sequence of coin flippings that produces one letter will not be the beginning part of the sequence of coin flippings that produces another letter. (2) For a random sequence produced by $p(x)$, the corresponding binary sequence is a sequence of iid realizations from Bernoulli(1/2), i.e., a sequence of random coin flippings. This is because the binary code is based on the coin flipping experiment that reproduces $p(x)$. So entropy(p) is the expected coding length. The unit of coding length is termed "bit."

This has a profound consequence. The random binary sequence cannot be further compressed in general. Otherwise, for example, for a random sequence of length n, if we can shorten it to a sequence of $.9n$, then there can only be $2^{.9n}$ sequences. That cannot cover all the 2^n sequences. So entropy(p) is the shortest expected coding length.

9.3 Kolmogorov complexity and randomness

A related concept is Kolmogorov complexity. For a binary sequence, its Kolmogorov complexity is the shortest binary machine code that produces this sequence.

For a binary sequence of length n, if we cannot further compressed it, then it must has all the statistical properties of a random binary sequence. For instance, the frequency of 1 or heads must be close 1/2. If it is 1/3, then we can focus on the subset of sequences with 1/3 heads. The number of such sequences is

$\binom{n}{n/3} \sim 2^{\rho n}$ for a $\rho < 1$. So we can number these sequences from 1, 2, ..., $2^{\rho n}$, and return the binary code of the order. This binary code has length less than ρn. If the frequency of head is close to 1/2, the number of such sequences is about 2^n according to the law of large numbers. So we cannot further compress it. Thus randomness = non-compressibility, and we can measure randomness by the shortest coding length.

9.4 Kullback-Leibler divergence

If we code the letter $X \sim p(x)$ by a wrong distribution $q(x)$ with code length $-\log_2 q(x)$ instead of $-\log_2 p(x)$, then the redundancy is

$$KL(p|q) = \mathbb{E}[-\log q(X)] - \mathbb{E}[-\log p(X)] = \mathbb{E}[\log(p(X)/q(X)],$$

which is the Kullback-Leibler divergence from p to q.

$$\mathbb{E}[q(X)/p(X)] = \sum_x [q(x)/p(x)]p(x) = \sum_x q(x) = 1.$$

According to Jensen inequality, for the concave function $\log(x)$, $\mathbb{E}[\log(X)] \leq \log(\mathbb{E}(X))$, so

$$\mathbb{E}[\log(q(X)/p(X))] \leq \log \mathbb{E}[q(X)/p(X)] = 0.$$

Thus $KL(p|q) \geq 0$.

In large deviation,

$$P(\bar{X}_n > t) \sim e^{-nKL(p_{\hat{\lambda}}|p)},$$

where $p_{\hat{\lambda}}$ is the tilted distribution so that $\mathbb{E}_{\hat{\lambda}}[X] = t$. This is the deepest place where we can find the Kullback-Leibler divergence.

10 Markov chain

10.1 Markov property

Suppose we make random walk on three states $S = \{1,2,3\}$. At each step, we randomly move to one of the other two states, regardless of the past history. Let X_t be the state at step t. Then $P(X_{t+1} = y|X_t = x, X_{t-1}, X_{t-2}, ..., X_0) = P(X_{t+1} = y|X_t = x) = K(x,y)$. That is, knowing the current state x, the probability that we end up in state y in the next step is determined, regardless where we come from.

A very important property of the Markov chain is $P(X_{t+1} = y) = \sum_x P(X_{t+1} = y, X_t = x) = \sum_x P(X_t = x)P(X_{t+1} = y|X_t = x) = \sum_x P(X_t = x)K(x,y)$. This is a direct application of the rule of total probability. $P(X_t = x)$ is called the marginal probability.

An intuitive way of understanding the above result is as follows. Consider a population of 1 million people move around the three states according to the random walk described above. We can interpret $P(X_t = x) \times 1$ *million* as the number of people in state x at time t. $P(X_{t+1} = y|X_t = x) = K(x,y)$ is the fraction of those in state x who will move to state y. Then $P(X_t = x)K(x,y)$ is the number of people who will move from x to y (i.e., those people who are in state x at time t and who will be in state y at time $t+1$). Thus $\sum_x P(X_t = x)K(x,y)$ will be the total number of people who will end up in y at time $t+1$, and this is $P(X_{t+1} = y) \times 1$ *million*. So you can visualize the change of the probability distribution of $(P(X_t = x), \forall x \in S)$ as the change of the distribution of the population over time t, as if we are watching a movie where each t is a frame of the movie. As $t \to \infty$, the distribution (either a probability distribution or a population distribution) will converge to a stationary distribution π, i.e., $P(X_t = x) \to \pi(x)$ for all $x \in S$, no matter what is the initial distribution or no matter where we start from.

In the above example, the conditional independence assumption enables us to conclude that among all the people in x at time t, the fraction of people who will move to y at time $t+1$ is $K(x,y)$ regardless of these people's past immigration histories.

The google pagerank algorithm assumes a random walk or Markov chain over the state space of all the webpages. The stationary distribution $\pi(x)$ can be interpreted as the proportion of web surfers who are on page x. So $\pi(x)$ is an indicator of the popularity of x.

10.2 Markov chain

The Markov property is such that given the present, the future is independent of the past. Let the state space be Ω, a Markov chain on Ω is determined by the transition probability

$$K(x,y) = P(X_{t+1} = y | X_t = x, X_{t-1}, ..., X_0) = P(X_{t+1} = y | X_t = x).$$

Let $p^{(t)}(x)$ be the marginal distribution of X_t. Then

$$\begin{aligned} p^{(t+1)}(y) &= P(X_{t+1} = y) \\ &= \sum_x P(X_{t+1} = y, X_t = x) \\ &= \sum_x P(X_{t+1} = y | X_t = x) P(X_t = x) \\ &= \sum_x p^{(t)}(x) K(x,y). \end{aligned}$$

Let K be the matrix $(K(x,y))$. Let $p^{(t)}$ be the row vector $(p^{(t)}(x))$. Then $p^{(t+1)} = p^{(t)}K$. By induction, $p^{(t)} = p^{(0)}K^t$.

Under very general conditions, $p^{(t)} \to \pi$, the stationary distribution, so that $\pi = \pi K$, i.e., $\pi(y) = \sum_x \pi(x)K(x,y)$.

The two-step transition

$$\begin{aligned} K^{(2)}(x,y) &= P(X_{t+2} = y | X_t = x) \\ &= \sum_z P(X_{t+2} = y, X_{t+1} = z | X_t = x) \\ &= \sum_z P(X_{t+2} = y | X_{t+1} = z, X_t = x) P(X_{t+1} = z | X_t = x) \\ &= \sum_z K(x,z)K(z,y) = K^2(x,y). \end{aligned}$$

In general, $K^{(t)} = K^t$.

10.3 Population migration

The way to visualize the Markov chain is to consider a population of say 1 million people moving around in the state space Ω. Then $p^{(t)}(x)$ can be considered the number of people in state x at time t. $p^{(t)}$ can be considered the distribution of the population at time t. π is the population distribution in the limit. $K(x,y)$ can be considered the fraction of the people in x who will go to y. The above calculations can then be interpreted in terms of numbers of people and fractions.

10.4 Reversibility or detailed balance

A special case is the reversible Markov chain where

$$\pi(x)K(x,y) = \pi(y)K(y,x)$$

for every (x,y). It is also called the detailed balance condition. If a chain is reversible with respect to π, then π is the stationary distribution, because

$$\sum_x \pi(x)K(x,y) = \sum_x \pi(y)K(y,x) = \pi(y)\sum_x K(y,x) = \pi(y).$$

The reversibility says that at stationarity, the number of people moving from x to y is the same as the number of people moving from y to x. If this is true for every pair of (x,y), then the numbers of people in different states will not change. That is, detailed balance implies overall balance.

10.5 Arrow of time and the second law of thermodynamics

Suppose 1 million people starts from a state, and then each person moves around according to the Markov transition probability K. Then the population starts to diffuse and eventually will reach a stationary distribution, so that the population distribution does not change beyond this point, even though the people are still moving around.

Let $KL(p^{(t)}|\pi)$ be the Kullback-Leibler divergence from $p^{(t)}$ to π. We will show that $KL(p^{(t)}|\pi) \geq KL(p^{(t+1)}|\pi)$. Let $p(x,y) = p^{(t)}(x)K(x,y)$ be the joint distribution of (X_t, X_{t+1}). Let $\pi(x,y) = \pi(x)K(x,y)$ be the joint distribution of (X_t, X_{t+1}) at stationarity. Let $p(y|x)$ and $\pi(y|x)$ be the conditional distribution of X_t given X_{t+1} under p and π respectively. In general $p(y|x) \neq \pi(y|x)$. $p(x,y) = p^{(t)}(x)K(x,y) = p^{(t+1}(y)p(x|y)$. $\pi(x,y) = \pi(x)K(x,y) = \pi(y)\pi(x|y)$. Thus

$$\begin{aligned}
KL(p|\pi) &= \mathbb{E}_p\left[\log \frac{p(X,Y)}{\pi(X,Y)}\right] \\
&= \mathbb{E}_p\left[\log \frac{p^{(t)}(X)K(X,Y)}{\pi(X)K(X,Y)}\right] \\
&= \mathbb{E}_p\left[\log \frac{p^t(X)}{\pi(X)}\right] = KL(p^{(t)}|\pi) \\
&= \mathbb{E}_p\left[\log \frac{p^{(t+1)}(Y)p(X|Y)}{\pi(Y)\pi(X|Y)}\right] \\
&= KL(p^{(t+1)}|\pi) + KL(p(x|y)|\pi(x|y)).
\end{aligned}$$

So $KL(p^{(t)}|\pi) \geq KL(p^{(t+1)}|\pi)$. Thus $p^{(t)}$ gets closer to π as t increases.

If π is uniform, then $\pi(x) = 1/|\Omega|$, where $|\Omega|$ is the number of states in Ω. Then $KL(p^{(t)}|\pi) = \mathbb{E}_{p^{(t)}}[\log p^{(t)} + \log|\mathscr{X}|] = -\text{entropy}(p^{(t)}) + \log|\Omega|$. In that case, the entropy of $p^{(t)}$ increases as t increases. This gives a statistical justification for the second law of thermodynamics.

10.6 Google pagerank

One notable application of Markov chain is the google pagerank algorithm, which is based on the random surfer model, i.e., with 85% probability, a surfer randomly click a link on the current page and move to the linked page. With 15% probability, the surfer randomly move to another page not necessarily linked by

the current page. Let K be the huge transition matrix (which we never explicitly record). Let $p^{(t)}$ be the marginal distribution, then $p^{(t)} \to \pi$.

The population interpretation is particularly relevant. $\pi(x)$ measures how many people in page x. The more people in a particular webpage at stationarity, the more popular this webpage is.

In order to calculate π, we can use $p^{(t+1)}(y) = \sum_x p^{(t)}(x)K(x,y)$. We may start $p^{(0)}$ to be the uniform distribution. After a number of steps, $p^{(t)}$ should be close to π.

There is a more direct way to think about it. If we want to know how popular a page y is, we may count how many other pages x are linked to it. But this has the shortcoming that those x themselves may not be very popular, i.e., not many pages z are linked to those x. So we also want to take into account the popularity of x themselves. This leads to the recursive relationship $p^{(t+1)}(y) = \sum_x p^{(t)}(x)K(x,y)$ and the fix point relationship $\pi(y) = \sum_x \pi(x)K(x,y)$, where $\pi(x)$ is the popularity of x.

10.7 Transition matrix: noun and verb

The transition matrix K contains the transition probabilities. This can be considered the noun meaning of K. There are also two verb meanings of K.

(1) Forward meaning as mixing. In $p^{(t+1)} = p^{(t)}K$, K acts on a row vector, and its verb meaning is mixing $p^{(t)}$ into $p^{(t+1)}$.

(2) Backward meaning as smoothing. We can also let K act on a column vector so that $g = Kh$. Both g and h can be considered functions defined on the state space Ω.

$$\begin{aligned} g(x) &= \sum_y K(x,y)h(y) \\ &= \sum_y h(X_{t+1} = y)P(X_{t+1} = y|X_t = x) \\ &= \mathbb{E}(h(X_{t+1})|X_t = x), \end{aligned}$$

which is local average of h around x. So the verb meaning of K is that it smoothes h into g. In general, if $g = K^{(s)}h$, then $g(x) = \mathbb{E}(h(X_{t+s})|X_t = x)$. Thus $K^{(s)}$ has the verb meaning as smoothing h for s steps, and changing h into g.

We can combine (1) and (2) above by attaching h to $p^{(t+1)} = p^{(t)}K$ to get $p^{(t+1)}h = p^{(t)}Kh$.

$$p^{(t+1)}h = \sum_x h(x)P(X_{t+1} = x) = \mathbb{E}[h(X_{t+1})].$$

Let $g = Kh$, then $g(x) = \mathbb{E}[h(X_{t+1}|X_t = x)]$ as explained before. $p^{(t)}g = \sum_x g(x)P(X_t = x) = \mathbb{E}[g(X_t)]$. Thus we have $\mathbb{E}[h(X_{t+1})] = \mathbb{E}[g(X_t)] = \mathbb{E}[h(X_{t+1}|X_t]]$. This is the Adam formula. It means the global averaging is the average of the local averaging.

10.8 Matrix eigenvalues, operator norm, and statistical underpinning

For the transition matrix K, we can write it as $K = V\Lambda U^\top$, with $U^\top = V^{-1}$, and Λ being diagonal. $KV = V\Lambda$, so the columns of V are the right eigenvectors of K. $U^\top K = \Lambda U^\top$, so the columns of U are the left eigenvectors of K. For the stochastic matrix K, the largest eigenvalue is 1, with the left eigenvector being π and the right eigenvector being $\mathbf{1}$, a vector of 1's. These can be seen from the verb meaning of K. $\pi K = \pi$, i.e., mixing π by K gives us π again. $K\mathbf{1} = \mathbf{1}$, i.e., smoothing a constant function gives us the same function again.

$K^{(t)} = V\Lambda^t U^\top \to \mathbf{1}\pi$, i.e., all the rows of $K^{(t)}$ approach π, because $\lambda_i^t \to 0$ for $i \neq 1$. The rate of convergence is determined by the second largest eigenvalue (in absolute value).

For a reversible K, we can let $S = \text{diag}(\pi)^{1/2} K \text{diag}(\pi)^{-1/2}$, then S is symmetric, and can be decomposed into $S = Q\Lambda Q^\top$. The square of the second largest eigenvalue is γ in the previous subsection.

We can also consider K as an operator in a Hilbert space whose inner product is defined by

$$\langle g, h \rangle_\pi = \sum_x g(x)h(x)\pi(x).$$

Then $\langle h, 1 \rangle_\pi = \mathbb{E}_\pi[h(X)]$, and $\|h\|_\pi^2 = \langle h, h \rangle_\pi = \mathbb{E}_\pi[h(X)^2]$. Let $g = Kh$, then $g(x) = \mathbb{E}[h(X_{t+1})|X_t = x]$. The largest eigenvalue is 1 with eigenvector $\mathbf{1}$. The second largest eigenvalue must have an eigenvector h such that $\langle h, 1 \rangle_\pi = \mathbb{E}_\pi[h(X)] = 0$, and the square of the second largest eigenvalue is

$$\gamma = \max_{h:\langle h,1\rangle_\pi = 0} \frac{\|Kh\|_\pi^2}{\|h\|_\pi^2} = \max_{h:\mathbb{E}_\pi[h(Y)]=0} \frac{\text{Var}[\mathbb{E}[h(Y)|X]]}{\text{Var}[h(Y)}.$$

10.9 Metropolis algorithm

The Metropolis algorithm is considered the number 1 algorithm in scientific computing by a panel of experts on computing. The algorithm is used to sample from a target distribution π. As an example, suppose there are three states $\Omega = \{1, 2, 3\}$. Suppose that the target distribution is $\pi(1) = .4$, $\pi(2) = .2$, and $\pi(3) = .4$. Suppose we run a base chain such that we randomly move to one of the other two states. The transition probability B of this base chain is

$$B = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

But this base chain will converge to the uniform distribution instead of the target distribution π. Consider the population migration perspective, suppose there are 1 million people, and the distribution of the population is already π, i.e., the numbers of people in the three states are .4 million, .2 million, and .4 million. If we follow the base chain, the population distribution will change. So we need to modify the base chain to maintain the distribution. According to the base chain, the number of people going from state 1 to state 2 is .2 million, and the number of people going from state 2 to state 1 is .1 million. We may allow all the .1 million people going from state 2 to state 1, but we shall only allow .1 million people going from state 1 to state 2 in order to main the balance, i.e., we need an extra randomization: for each person going from state 1 to state 2, he or she will be allowed to go through only with 1/2 probability. The same for other pairs of states.

More generally, we start from a base chain $B(x, y)$. Suppose $X_t = x$. We then propose to move to y by sampling y from $B(x, y)$. We accept this proposal with probability $\min[1, (\pi(y)B(y,x))/(\pi(x)B(x,y))]$. If the proposal is accepted, then $X_{t+1} = y$. Otherwise, $X_{t+1} = x$.

The intuition behind this algorithm is as follows. Suppose a population of 1 million people move around in the state space. If the population distribution is already π, then we want the distribution to remain π after the Markov transition. Now consider two states x and y. According to $B(x, y)$, the number of people who propose to move from x to y is $\pi(x)B(x, y)$. The number of people who propose to move from y to x is $\pi(y)B(y, x)$. Suppose $\pi(x)B(x, y) \geq \pi(y)B(y, x)$. Then we can allow all the $\pi(y)B(y, x)$ people who propose to move from y to x to make the move. But for those $\pi(x)B(x, y)$ people who propose to move from x to y, we can only allow $\pi(y)B(y, x)$ people to make the move in order to maintain the detailed balance. That is, for each person who proposes such a move, we allow this person to make the move with probability $(\pi(y)B(y,x))/(\pi(x)B(x,y))$. This will maintain the detailed balance between every two states, and thus it will maintain the stationarity. It is like we have a visa office to decide whether to grant a person the visa to go through.

The transition probability for the resulting Markov chain is

$$M(x,y) = B(x,y)\min\left[1, \frac{\pi(y)B(y,x)}{\pi(x)B(x,y)}\right],$$

for $x \neq y$. $M(x,x) = 1 - \sum_{y \neq x} M(x,y)$.

Then $\pi(x)M(x,y) = \min(\pi(y)B(y,x), \pi(x)B(x,y)) = \pi(y)M(y,x)$ for $x \neq y$.

For the example above,

$$M = \begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/4 & 1/4 \end{bmatrix}$$

10.10 Gibbs sampler

The Gibbs sampler is designed to sample from multi-dimensional distribution $p(x_1, ..., x_d)$. The algorithm is very simple. In each iteration, for $i = 1, ..., d$, we sample $X_i \sim p(x_i | x_{-i})$, where the notation x_{-i} denotes the *current* values of all the other components except i. Such a move preserves the target distribution. Suppose $X \sim p(x)$, $x = (x_1, ..., x_d)$. Then $p(x) = p(x_{-i})p(x|x_{-i})$. In stationarity, $X_{-i} \sim p(x_{-i})$, and $[X_i | X_{-i}] \sim p(x_i|x_{-i})$. If we change X_i by another independent copy from the same distribution, we are not going to change the joint distribution.

As an example, consider sampling from a bivariate normal distribution $p(x,y)$ with correlation ρ. Let (X_t, Y_t) be the current values of (X,Y), then we sample $X_{t+1} = \rho Y_t + \varepsilon_1$, where $\varepsilon_1 \sim \mathrm{N}(0, 1-\rho^2)$. After that we sample $Y_{t+1} = \rho X_{t+1} + \varepsilon_2$, where $\varepsilon_2 \sim \mathrm{N}(0, 1-\rho^2)$, and both ε_1 and ε_2 are independent of anything else.

We can visualize the Gibbs sampler for a special case where the target distribution is a uniform distribution over a two-dimensional shape (such as a big island). Suppose one million people start from a certain point. Then at each iteration, we let people randomly relocate horizontally and then vertically. Then people will start to diffuse, and gradually will reach the uniform distribution on the island. At this point, horizontal and vertical re-shuffling of people will not change the population distribution.

10.11 Markov random field, undirected graphical model

In ond-dimension, the Ising model for $(X_1, ..., X_i, ..., X_n)$ is

$$p(x_1, ..., x_i, ..., x_n) = \frac{1}{Z(\beta)} \exp\left[\sum_{i=0}^{n} \beta x_i x_{i+1}\right],$$

where $x_i \in \{+1, -1\}$ means whether x_i spins up or down, and we fix x_0 and x_{n+1} as the boundary condition. $\beta > 0$ encourages neighboring x_i to have the same spin. If β depends on i, then the model is called spin glass model. $Z(\beta)$ is the normalizing constant to make p sum to 1. In statistical physics, β is called the inverse temperature, and $-\sum_i x_i x_{i+1}$ is called energy. The probability distribution prefers low energy states.

Define $x_{-i} = (x_j, j \neq i)$, we can show that

$$p(x_i \mid x_{-i}) = p(x_i | x_{i-1}, x_{i+1})$$

for $i = 1, ..., n$.

The model can be generalized to two dimensional or three dimensional. It is an undirected graphical model because it cannot be factorized in the way as in the Bayes net. In two or three dimensions, the Ising model exhibits phase transition. That is, the distribution can be different with different boundary conditions.

The Ising model is a special case of a Markov random field model, where $p(x_i \mid x_{-i})$ only depends on the neighbors of x_i.

We have seen such a distribution in our study of large deviation rate before.

11 Continuous time processes

11.1 Markov jump process and transition rate

For one period model, we may assume that

$$P(X_{t+\Delta t} = j | X_t = i) = a_{ij}\Delta t, \text{ for } i \neq j.$$

Then

$$P(X_{t+\Delta t} = i | X_t = i) = 1 - \sum_{j \neq i} a_{ij}\Delta t = 1 + a_{ii}\Delta t,$$

where we define

$$a_{ii} = -\sum_{j \neq i} a_{ij}.$$

Here for more direction connection to matrix, we use i and j instead of x and y for indices.

Let $K^{(\Delta t)}$ be the one-period transition matrix, then

$$K^{(\Delta t)} = I + A\Delta t,$$

where $A = (a_{ij})$. Thus

$$K^{(t)} = (K^{(\Delta t)})^{t/\Delta t} = (1 + A\Delta t)^{t/\Delta t} \to \exp(At),$$

where for a matrix M, $\exp(M) = \sum_{k=0}^{\infty} M^k/k!$. $K(t)$ forms a semi-group (i.e., $K(t)K(s) = K(t+s)$, $K(0) = I$, but there is no inverse $K(t)^{-1}$), with A being the generator.

11.2 Forward and backward equations of jump process

We can derive the forward and backward equations by the last step and first step analyses respectively.

For forward equation,

$$K^{(t+\Delta t)} = K^{(t)}K^{\Delta t} = K^{(t)}(I + A\Delta t),$$

thus

$$(K^{(t+\Delta t)} - K^{(t)})/\Delta t = K^{(t)}A,$$

so as $\Delta t \to 0$,

$$dK^{(t)}/dt = K^{(t)}A.$$

For backward equation,

$$K^{(t+\Delta t)} = K^{(\Delta t)}K^{(t)} = (1 + A\Delta t)K^{(t)},$$

thus

$$(K^{(t+\Delta t)} - K^{(t)})/\Delta t = AK^{(t)},$$

so

$$dK^{(t)}/dt = AK^{(t)}.$$

11.3 Brownian motion, $\sqrt{\Delta t}$ notation, second order Taylor expansion

The Brownian motion is defined by

$$X_{t+\Delta t} = X_t + \sigma \varepsilon_t \sqrt{\Delta t},$$

where $\mathbb{E}(\varepsilon_t) = 0$ and $\mathrm{Var}(\varepsilon_t) = 1$. Let $X_0 = x$. If we divide the interval $[0,t]$ into n periods, so that $\Delta t = t/n$. Write ε_t within each period as ε_i for $i = 1, ..., n$. Then

$$X_t = x + \sum_{i=1}^{n} \sigma \varepsilon_i \sqrt{\frac{t}{n}} = x + \sigma \sqrt{t} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \to \mathrm{N}(x, \sigma^2 t)$$

according to the central limit theorem. The above equation also shows why we need $\sqrt{\Delta t}$ scaling in order to make the $\mathrm{Var}(X_t)$ to be independent of n or Δt. This $\sqrt{\Delta}$ scaling was first noticed by Einstein in his 1905 analysis of Brownian motion. Einstein explained that the velocity $dX(t)/dt = (X_{t+\Delta t} - X_t)/\Delta t = \sigma \varepsilon_t / \sqrt{\Delta t} \to \infty$.

The $\sqrt{\Delta t}$ notation also makes it clear that we should use second order Taylor expansion in our analysis. The second order Taylor expansion underlies much of the calculations in Brownian motion, including the Ito calculus.

11.4 Generator: noun and verb

The generator A in the Markov jump process is the rate of transition, $K(\Delta t) = I + A\Delta t$, so

$$A = \frac{K(\Delta t) - I}{\Delta t} = \frac{dK(t)}{dt}|_{t=0},$$

since $K(0) = I$. A is the rate of transition.

Now consider the verb meaning of K. For forward equation, $dK^{(t)}/dt = K^{(t)}A$, $p^{(0)}K^{(t)} = p^{(t)}$. So $dp^{(t)}/dt = p^{(t)}A$. For backward equation, $dK^{(t)}/dt = AK^{(t)}$, $g^{(t)} = K^{(t)}h$ is to smooth h for a period of time t. So $dg^{(t)}/dt = Ag^{(t)}$, with $g^{(0)} = h$.

The verb meaning of A in the backward context is

$$g = Ah = (K^{(\Delta t)} - I)h/\Delta t,$$

so

$$g(x) = [\mathbb{E}(h(X_{t+\Delta t})|X_t = x]) - h(x)]/\Delta t.$$

The generator of the Brownian motion has a verb meaning. We can let A act on a testing function h, and let $g = Ah$, then

$$\begin{aligned} g(x) &= \frac{\mathbb{E}[h(X_{t+\Delta t})|X_t = x] - h(x)}{\Delta t} \\ &= \frac{\mathbb{E}[h(x + \sigma \varepsilon_t \sqrt{\Delta}) - h(x)}{\Delta t} \\ &= \frac{\mathbb{E}[h(x) + h'(x)\sigma \varepsilon_t \sqrt{\Delta} + \frac{1}{2} h''(x) \sigma^2 \varepsilon_t^2 \Delta t + o(\Delta t)] - h(x)}{\Delta t} \\ &= \frac{\sigma^2}{2} h''(x). \end{aligned}$$

Thus $A = \frac{\sigma^2}{2} \frac{\partial^2}{\partial x^2}$, which is an operator (indeed a verb).

This can be seen more clearly if we consider a special case where $\varepsilon_t = +1$ with probability 1/2, and $\varepsilon_t = -1$ with probability 1/2. Let $X_0 = 0$. Then X_t must be a multiple of $\Delta x = \sigma \sqrt{\Delta t}$. So the state space is

also discretized. Then we can write $K^{(\Delta t)}$ as an infinite dimensional matrix, so that $K^{(\Delta t)}(x,x+\Delta x) = 1/2$ and $K^{(\Delta t)}(x,x-\Delta x) = 1/2$. Then the matrix $A = (K^{(\Delta t)} - I)/\Delta t = \frac{\sigma^2}{2\Delta x^2} D$, where the matrix D is such that $D(x,x+\Delta x) = D(x,x-\Delta x) = 1$ and $D(x,x) = -2$. As $\Delta x \to 0$, A becomes the second derivative operator.

The following is what the matrix D looks like.

$$\begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}$$

If we multiple it to a vector, it is like taking the second derivative. This matrix view makes clear both the noun meaning and verb meaning of A.

11.5 Heat equations for Brownian motion

Let $[X_t|X_0 = x] \sim K(t,x,y)$, i.e., $K(t,x,y)$ as a function of y is the density of X_t given $X_0 = x$. Then the backward equation $\partial K^{(t)}/\partial t = AK^{(t)}$ is

$$\frac{\partial K(t,x,y)}{\partial t} = \frac{\sigma^2}{2} \frac{\partial K(t,x,y)}{\partial x^2}.$$

The forward equation $\partial K^{(t)}/\partial t = K^{(t)}A$ is

$$\frac{\partial K(t,x,y)}{\partial t} = \frac{\sigma^2}{2} \frac{\partial K(t,x,y)}{\partial y^2}.$$

Both are heat equations. Einstein derived the forward equation and solved it to get the distribution of $X_t \sim \mathrm{N}(0,\sigma^2 t)$. We can also get the distribution of X_t by central limit theorem as above. Both the heat equation and the central limit theorem can be dealt with by the Fourier transform.

The forward equation is a bit more complicated than the backward equation, because we need to understand the meaning of KA. If we think about A as an infinite matrix as in the above subsection, we can understand why it is the second derivative with respect to y.

From an operator point of view, we can let it act on a testing function h. Assume $\sigma^2/2 = 1$ for simplicity. Applying integral by parts (and assuming $h(y)$ goes to 0 sufficiently fast as $y \to \pm\infty$),

$$\begin{aligned} K^{(t)}Ah &= \int K(t,x,y)h''(y)dy \\ &= -\int \frac{\partial}{\partial y} K(t,x,y)h'(y)dy \\ &= \int \frac{\partial^2}{\partial y^2} K(t,x,y)h(y)dy, \end{aligned}$$

so $K^{(t)}A = \frac{\partial^2}{\partial y^2} K(t,x,y)$.

The forward meaning of A is actually the transpose of the backward meaning of A.

We can derive the heat equations directly by first step analysis and last step analysis. For example, in

the last step analysis, let $X_{t+\Delta t} = X_t + \sigma\varepsilon\sqrt{\Delta t}$, then

$$\begin{aligned} K(x,y,t+\Delta t) &= \int K(x,y-\sigma\varepsilon\sqrt{\Delta t},t)f(\varepsilon)d\varepsilon \\ &= \int (K(x,y,t) - \frac{\partial}{\partial y}K(x,y,t)\sigma\varepsilon\sqrt{\Delta t} + \frac{\sigma^2}{2}\frac{\partial^2}{\partial y^2}K(x,y,t)\varepsilon^2\Delta t)f(\varepsilon)d\varepsilon \\ &= K(x,y,t) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial y^2}K(x,y,t)\Delta t, \end{aligned}$$

which leads to the forward equation. This was how Einstein derived the equation.

11.6 Fokker-Planck

For stochastic differential equation, the forward equation is also called the Fokker-Planck equation.

Let $p_t(x)$ be the probability distribution of X_t. We can imagine a million particles moving independently according to SDE (e.g., a drop of milk in your coffee), p_t is the distribution of these particles over time. We want to know how p_t changes over time.

Let $h(x)$ be a test function within a certain function class $\mathscr{C}$. For a probability distribution $p(x)$, $\mathbb{E}_p[h(X)] = \int h(x)p(x)dx$. If two distributions p and q agree with $\mathbb{E}_p[h(X)] = \mathbb{E}_q[h(X)]$ for all $h \in \mathscr{C}$, then $p = q$.

Second order Taylor:

$$\begin{aligned} \mathbb{E}[h(X_{t+\Delta t})] &= \mathbb{E}[h(X_t + \mu(X_t)\Delta t + \sigma(X_t)\sqrt{\Delta t}Z_t)] \\ &= \mathbb{E}[h(X_t)] + \mathbb{E}[h'(X_t)\mu(X_t)]\Delta t + \frac{1}{2}\mathbb{E}[h''(X_t)\sigma(X_t)^2]\Delta t + o(\Delta t). \end{aligned}$$

$$\frac{\mathbb{E}[h(X_{t+\Delta t})] - \mathbb{E}[h(X_t)]}{\Delta t} \to \mathbb{E}[h'(X_t)\mu(X_t)] + \frac{1}{2}\mathbb{E}[h''(X_t)\sigma(X_t)^2].$$

The left hand side

$$\begin{aligned} \frac{\mathbb{E}[h(X_{t+\Delta t})] - \mathbb{E}[h(X_t)]}{\Delta t} &= \int \frac{p_{t+\Delta t}(x) - p_t(x)}{\Delta t}h(x)dx \\ &\to \int \frac{\partial p_t(x)}{\partial t}h(x). \end{aligned}$$

Integral by part (assuming tails go to zero):

$$\begin{aligned} \int \frac{\partial p_t(x)}{\partial t}h(x)dx &= \int h'(x)\mu(x)p_t(x)dx + \frac{1}{2}\int h''(x)\sigma(x)^2 p_t(x)dx \\ &= -\int \frac{\partial}{\partial x}[\mu(x)p_t(x)]h(x)dx - \int \frac{\partial}{\partial x}[D(x)p_t(x)]h'(x)dx \\ &= -\int \frac{\partial}{\partial x}[\mu(x)p_t(x)]h(x)dx + \frac{\partial^2}{\partial x^2}[D(x)p_t(x)]h(x)dx \end{aligned}$$

where $D(x) = \sigma(x)^2/2$. Thus we have Fokker-Planck partial differential equation:

$$\frac{\partial p_t(x)}{\partial t} = -\frac{\partial}{\partial x}[\mu(x)p_t(x)] + \frac{\partial^2}{\partial x^2}[D(x)p_t(x)].$$

Consider the situation where $\mu(x) = 0$, and $\sigma(x)$ is constant. Then X_t is Brownian motion. The distribution of the 1 million particles tend to diffuse from high density regions to low density regions. Thus if

the curvature of the density function at x is positive, $p_t(x)$ will decrease. Otherwise it will increase. The resulting PDE is the heat equation first derived by Einstein.

Now consider the situation where $\sigma(x) = 0$ and $\mu(x)$ is a positive constant, i.e., the distribution moves to the right. If the slope of $p_t(x)$ is positive, then the density at x is higher than the density at $x - \Delta x$, which will move to x, thus $p_t(x)$ will decrease.

11.7 Langevin

Let the target density be the Gibbs distribution

$$\pi(x) = \frac{1}{Z} \exp(-U(x)),$$

where $U(x)$ is the energy function. Any distribution can be written in such a form. The Langevin dynamics[**?**] is

$$X_{t+\Delta t} = X_t - \frac{1}{2} U'(X_t)\Delta t + \sqrt{\Delta t} Z_t,$$

with $\mu(x) = -\frac{1}{2}U'(x)$, $\sigma(x) = 1$. The Fokker-Planck is

$$\frac{\partial p_t(x)}{\partial t} = -\frac{\partial}{\partial x}[U'(x)p_t(x)/2] + \frac{\partial^2}{\partial x^2}[p_t(x)/2].$$

Since

$$\frac{\partial}{\partial x}\pi(x) = -U'(x)\pi(x),$$

$\pi(x)$ is the solution to $\partial p_t(x)/\partial t = 0$. Thus π is the stationary distribution of Langevin.

For Langevin or Markov process in general, $KL(p_t|\pi) \to 0$ monotonically, which underlies the second law of thermodynamics.

$$KL(p_t|\pi) = -\text{entropy}(p_t) + \mathbb{E}_{p_t}[U(X_t)] + \log Z.$$

The Langevin has the gradient descent component $X_{t+\Delta t} = X_t - U'(X_t)$, which decreases the energy $U(X_t)$, and moves X_t toward the local mode. It also has a diffusion component $\sqrt{\Delta t} Z_t$, which increases the entropy of p_t, and may help X_t jump out of the local modes.

One may also consider Langevin as doing a variational approximation of π by p_t.

11.8 Simulated annealing

The Gibbs distribution also has a temperature term T,

$$\pi_T(x) = \frac{1}{Z_T} \exp\left(-\frac{U(x)}{T}\right).$$

At high temperature, the distribution is diffused. At low temperature, the distribution focuses on the local modes. At the temperature approaches zero, the distribution focuses on the global mode.

The Langevin is

$$X_{t+\Delta t} = X_t - \frac{1}{2T} U'(X_t)\Delta t + \sqrt{\Delta t} Z_t.$$

At high temperature, Langevin is diffusion. At low temperature, Langevin is gradient descent.

The process of simulated annealing is to reduce the temperature slowly, in order to find the global minimum of $U(x)$. This is the process of making strong metal.

11.9 Geometric Brownian motion

The geometric Brownian motion is often used as a model of stock price,

$$X_{t+\Delta t} = X_t + \mu \Delta t X_t + \sigma \varepsilon_t \sqrt{\Delta t} X_t.$$

This stochastic difference equation becomes the stochastic differential equation if we let $\Delta t \to 0$, i.e.,

$$dX_t = \mu X_t dt + \sigma X_t dB_t,$$

where B_t is the continuous version of the Brownian motion

$$B_{t+\Delta t} = B_t + \varepsilon_t \sqrt{\Delta t}.$$

Letting $\Delta t = t/n$, we can solve the stochastic difference equation by the second order Taylor expansion of

$$\log(1+x) = x - \frac{x^2}{2} + o(x^2),$$

with the central limit theorem

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \to Z \sim \mathrm{N}(0,1),$$

and the law of large numbers

$$\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \to 1.$$

More specifically,

$$\begin{aligned} \log X_{t+\Delta t} &= \log X_t + \log(1 + \mu \Delta t + \sigma \varepsilon_i \sqrt{\Delta t}) \\ &= \log X_t + \mu \Delta t + \sigma \varepsilon_i \sqrt{\Delta t} - \frac{1}{2} \sigma^2 \varepsilon_i^2 \Delta t \end{aligned}$$

Thus

$$\begin{aligned} \log X_t &= \log X_0 + \sum_{i=1}^{n} \mu \frac{t}{n} + \sum_{i=1}^{n} \sigma \varepsilon_i \sqrt{\frac{t}{n}} - \frac{1}{2} \sigma^2 \sum_{i=1}^{n} \varepsilon_i^2 \frac{t}{n} \\ &= \log X_0 + \mu t + \sigma \sqrt{t} Z - \frac{\sigma^2 t}{2} \end{aligned}$$

So $X_t = X_0 \exp[\mu t - \sigma^2 t/2 + \sigma \sqrt{t} Z]$.

11.10 Ito calculus

The second order Taylor expansion and the law of large number lead to $\sum_{i=1}^{n} \varepsilon_i^2 \frac{t}{n} \to t$. This underlies the Ito calculus that says that $dB_t^2 = \varepsilon_t^2 dt = dt$ as far as the stochastic integral is concerned. We you Taylor expand a function that involves dB_t, you can replace dB_t^2 in the second order Taylor expansion by dt.

11.11 Martingale

Let $\varepsilon_t, t = 1,2,...$ be a sequence of random variables that drive the process, and let $X_t = f_t(\varepsilon_1,...,\varepsilon_t)$. Let $\mathscr{F}_t = \sigma(\varepsilon_1,...,\varepsilon_t)$, i.e., the σ-algebra generated by $\varepsilon_1,...,\varepsilon_t$, i.e., all the meaningful statements about the history up to time t. We say that X_t is adapted to $\mathscr{F}_t$. The latter is called filtration, i.e., the unfolding of history.

The sequence $X_t, t = 1,2,...$ is a martingale if for any $s < t$, $\mathbb{E}(X_t|\mathscr{F}_s) = X_s$. This is a notion of no drift or fair game. Let $Y_t = X_{t+1} - X_t$, then $\mathbb{E}[Y_t|\mathscr{F}_t] = 0$, i.e., the expected drift or the change of fortune is 0, and $X_t = X_0 + \sum_{i=1}^t Y_i$.

The above definition also holds for continuous time processes.

Besides Markov chain, Martingale is another way to depart from iid sequences. While Markov chain is for concrete modeling, Martingale is a theoretical structure that has many useful properties, including some properties that are true for iid sequences. This makes it a powerful tool for theoretical analysis of dependent sequences and stochastic processes.

11.12 Conditional expectation as anticipation

In both Markov chain and martingale in particular, we make heavy use of conditional expectation. Here the word "expectation" is really appropriate, because the original meaning of "expectation" is something like anticipation, i.e., what we can expect in the future given the current situation. In martingale, we have $\mathbb{E}[X_t|\mathscr{F}_s)$ for $s < t$. It means our anticipation of the future X_t given what has happened up to $s < t$. In Markov chain, we often need to compute $\mathbb{E}[h(X_t)|X_s = x] = \sum_y P(X_t = y|X_s = s)h(y) = K^{(t-s)}h$ as a smoothing of h for the period $[s,t]$. This is again an anticipation of $h(X_t)$ given X_s. Because of the Markov property $[X_t|X_s] = [X_t|\mathscr{F}_s]$, so $\mathbb{E}[h(X_t)|X_s] = \mathbb{E}[h(X_t)|\mathscr{F}_s]$.

For the Markov process X_t, let $V_t = \mathbb{E}[h(X_T)|X_t]$. For $s < t$, we have

$$\mathbb{E}[V_t|\mathscr{F}_s] = \mathbb{E}[V_t|X_s] = \mathbb{E}[\mathbb{E}[h(X_T)|X_t, X_s] = \mathbb{E}[h(X_T)|X_s] = V_s,$$

so V_t form a martingale with respect to the ε_t that drives X_t. That is, our anticipation of the same quantity $h(X_T)$ over time as we accumulate more and more information is a martingale.

Intuitively, you can imagine starting from X_s, there can be M trajectories if we simulate the Markov process, so at time t, there can be M of X_t's. Starting from each X_t, there can be N trajectories branching out from the current trajectory, until we reach time T. Then there are MN X_T's. The average of these MN $h(X_T)$'s gives us $\mathbb{E}[h(X_T)|X_s]$. We can first average the N $h(X_T)$'s on each of the M beaches to get $V_t = \mathbb{E}[h(X_T)|X_t]$. Then average the M V_t's to get V_s, so $\mathbb{E}[V_t|\mathscr{F}_s] = V_s$. That is, the overall anticipation of $h(X_T)$ at time s is the anticipation of the anticipation at time t.

11.13 Risk neutral expectation

Consider the Black-Sholes formula for option pricing. Suppose the stock price X_t follows the geometric Brownian motion P: $X_{t+\Delta t} = X_t + aX_t\Delta t + \sigma X_t\varepsilon_t\sqrt{\Delta t}$. One share of option is such that you have the right but not the obligation to buy one share of stock at the strike price K at the maturation time T, i.e., the value at time T is $h(X_T) = \max(0, X_T - K)$. If the price of the stock at time t is x, then the value of one share of option is

$$V(x,t) = e^{-r(T-t)}\mathbb{E}_Q[h(X_T)|X_t = x],$$

where r is the interest rate, and Q is the geometric Brownian motion $X_{t+\Delta t} = X_t + rX_t\Delta t + \sigma X_t\varepsilon_t\sqrt{\Delta t}$. Q is called the risk neutral measure. $V(x,t)$ is the expected time-discounted value of the option assuming an imaginary risk neural process.

For this imaginary risk neural process Q, the time discounted $\tilde{V}_t = e^{-t}V(X_t,t)$ form a martingale with respect to $\{\varepsilon_t\}$, according to our discussion in the previous subsection.