A Note on Coin Flipping

Ying Nian Wu
UCLA Department of Statistics
Fall 2008, while teaching STAT 200A

1 Why coin flipping?

For all its simplicity, a study of coin flipping reveals many of the deep results and useful intuitions in probability theory. In particular, we shall see the following things in our adventure in coin flipping:

(1) Weak and strong laws of large number.
(2) Central limit theorem.
(3) Large deviation.
(4) Probability distributions: Bernoulli, Binomial, Poisson, Normal and Exponential.
(5) Expectation and variance.
(6) Stochastic processes: random walk, Brownian motion, Poisson process.
(7) Coding and complexity: entropy, Kullback-Leibler divergence, equipartition.

The emphasis is on how to think about various concepts. Of course, we are also going to do some calculations.

2 Sample space and events

Suppose we flip a coin n times. Let Ω be the set of all possible sequences. Ω is called sample space. $|\Omega| = 2^n$, where $|\Omega|$ denotes the number of elements in Ω.

An event is a statement about the outcome. The event is represented by the subset of all the outcomes that satisfy this statement. For instance, let A be the event that the number of heads is k, where k is an integer from 0 to n. Then A is the set of all the sequences that have k heads.

3 Random variables

For each sequence $\omega \in \Omega$, we can represent it by a sequence of 0s and 1s, $Z_1(\omega)$, ..., $Z_n(\omega)$, where $Z_i(\omega) = 1$ if the i-th flip of ω is head, and $Z_i(\omega) = 0$ otherwise. Let $X(\omega)$ be the number of heads in sequence ω. Then $X(\omega) = Z_1(\omega) + ... + Z_n(\omega)$.

For a random sequence ω, the corresponding $Z_i(\omega)$ is a binary random number for each i, and $X(\omega)$ is a random number whose value is between 0 and n. We call Z_i and X random variables.

The random variables are actually functions. For example, the random variable X is a function that maps Ω to a set of numbers $\{0, 1, ..., n\}$, so that each outcome ω is labeled by a number $X(\omega) \in \{0, 1, ..., n\}$.

Let A be the event that the number of heads is k, then $A = \{\omega : X(\omega) = k\}$, i.e., the set of all the sequences whose numbers of heads are exactly k. In general, an event can be expressed as an equation or an inequality in terms of the random variables. So the introduction of random variables is mostly a matter of language. Instead of using English, we use equations and inequalities to describe the outcome.

4 Equally likely setting

If the coin is fair and the flips are independent, then all the 2^n sequences in Ω are equally likely. So we can imagine that instead of flipping the fair coin independently n times to get the sequence, we simply randomly pick a sequence from the population of these 2^n sequences.

5 Number of heads

Now let us study the random variable X, which is the number of heads if we flip the fair coin n times independently. We want to calculate $P(X = k)$ for $k = 0, 1, ..., n$.

Based on the consideration in Section 4, we only need to know how many sequences have exactly k heads. Let A_k be the set of sequences with k heads. Then $P(X = K) = P(A_k) = |A_k|/|\Omega| = |A_k|/2^n$, where $|A_k|$ is the size of A_k, i.e., the number of sequences in A_k.

6 Population and classes

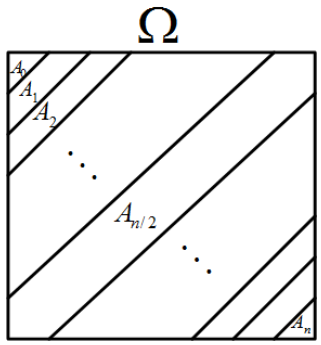

Figure 1: Partition population into classes.

We can imagine that Ω is a population, which can be partitioned into $n + 1$ classes, A_0, A_1, ..., A_n. All the members in the same class share the same number of heads, i.e., if $\omega \in A_k$, then $X(\omega) = k$. See Figure 1 for an illustration.

7 The size of a class

Now we want to measure the size of A_k. We can show that $|A_k| = \binom{n}{k}$.

The reason is as follows. In order to produce a sequence with k heads, we only need to choose k flips from the sequence of n flips, and let these k flips be heads, and the rest of the $n - k$ flips be tails. The number of different ways of choosing k flips from the n flips is $\binom{n}{k}$, and this is the number of sequences in A_k.

$\binom{n}{k}$ is given by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

8 How large is $\binom{n}{k}$

For large n, we want to have a simpler expression for $|A_k| = \binom{n}{k}$. In order to do that, we need the Stirling formula, which says that

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

The meaning of this formula is that the ratio between the left hand side and the right hand side goes to 1 as n goes to infinity.

Let $\alpha = k/n$, which is the frequency of heads, then for large n,

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{(n\alpha)!(n(1-\alpha))!} \\
&\sim \frac{\sqrt{2\pi n} n^n e^{-n}}{[\sqrt{2\pi n\alpha}(n\alpha)^{n\alpha} e^{-n\alpha}][\sqrt{2\pi n(1-\alpha)}(n(1-\alpha))^{n(1-\alpha)} e^{-n(1-\alpha)}]} \\
&= \frac{1}{\sqrt{2\pi n\alpha(1-\alpha)}\alpha^{n\alpha}(1-\alpha)^{n(1-\alpha)}} \\
&= C(n,\alpha) 2^{nh(\alpha)},
\end{aligned}$$

where

$$\begin{aligned}
C(n,\alpha) &= \frac{1}{\sqrt{2\pi n\alpha(1-\alpha)}}, \\
h(\alpha) &= -[\alpha \log_2 \alpha + (1-\alpha)\log_2(1-\alpha)].
\end{aligned}$$

We write the above result as $\binom{n}{k} \sim C(n,\alpha)2^{nh(\alpha)}$, meaning that the ratio between the left hand side the the right hand side goes to 1 as n goes to infinity. In taking the limit, we hold α fixed.

9 Entropy function

The function $h(\alpha)$ is a measure of entropy. Later we will make this clear. But for now, we only need to see what this function looks like.

$h(\alpha)$ is defined for $\alpha \in (0,1)$. $h'(\alpha) = -\log\alpha + \log(1-\alpha)$, and $h''(\alpha) = -1/\alpha - 1/(1-\alpha) < 0$. So $h(\alpha)$ is a concave function, and it achieves its maximum at $\alpha = 1/2$, where $h(1/2) = 1$, and $h'(1/2) = 0$.

10 Class sizes for large n

The results in Sections 8 and 9 have deep implications. If α is away from $1/2$, say, $|\alpha - 1/2| > \epsilon$ for $\epsilon > 0$, then the size of the class A_k or $A_{n\alpha}$, which is $\binom{n}{k} \sim C(n,\alpha)2^{nh(\alpha)}$, will becomes smaller and smaller relative to the size of the whole population $|\Omega| = 2^n$. In other words, the proportion of the class $A_{n\alpha}$, which is $C(n,\alpha)2^{nh(\alpha)}/2^n$, goes to 0 at an exponential rate.

Because of the shape of the entropy function $h(\alpha)$, if $|\alpha - 1/2| > \epsilon$, then $h(\alpha) < h(1/2+\epsilon) = h(1/2-\epsilon) = r < 1$. So the size of class $A_{n\alpha}$ is smaller than $C(n,\alpha)2^{nr}$, and the proportion of class $A_{n\alpha}$ is smaller than $C(n,\alpha)2^{-n(1-r)}$, which is exponentially small.

11 Sum of exponentials

Even if we put together all those classes A_k with $|k/n - 1/2| > \epsilon$, their total size is still just something like $C2^{-n(1-r)}$, where C is a number that changes slowly compared to the exponential term.

To see this fact, consider a simpler example. Suppose we have one term $2^{-.5n}$, and we have another term $2^{-.3n}$. If we put them together, the sum behaves like $2^{-.3n}$, because $2^{-.5n}$ is negligibly small compared to $2^{-.3n}$ when n is large. That is, if we sum up a limited number of terms that goes to 0 exponentially fast, the sum behaves like the largest term.

So the total proportion of those classes A_k with $|k/n - 1/2| > \epsilon$ goes to 0 at an exponential rate. Or in other words, the total proportion of those classes A_k with $|k/n - 1/2| \leq \epsilon$ goes to 1, no matter how small ϵ is.

So as far as frequency of heads is concerned, 1/2 is the typical value, in the sense that nearly all of the 2^n sequences share this value (with small deviations no more than ϵ).

12 Weak law of large number

Let $A = \{\omega : |X(\omega)/n - 1/2| > \epsilon\}$. Then A is the union of those classes A_k with $|k/n - 1/2| > \epsilon$, and $|A|/|\Omega|$ goes to 0 at an exponential rate.

Translating this into probability, $P(A) = |A|/|\Omega|$ goes to 0, i.e.,

$$P(|\frac{X}{n} - \frac{1}{2}| > \epsilon) \to 0.$$

This is the so called weak law of large number. It says that if we flip a fair coin a large number of times, then the frequency of heads is close to 1/2.

13 Large deviation

More precisely, because the proportion of A is $C2^{-n(1-r)}$ according to Section 10,

$$\frac{1}{n} \log_2 P(|\frac{X}{n} - \frac{1}{2}| > \epsilon) \to -(1-r) < 0.$$

Such a result is called a large deviation result, because if $|X/n - 1/2| > \epsilon$, $|X - n/2| > \epsilon n$, i.e., we look at those sequences ω whose number of heads $X(\omega)$ deviates from $n/2$ by at least ϵn. No matter how small ϵ is, as we increase n, this deviation will become large.

14 Population histogram, probability distribution, and tail probability

For all the 2^n numbers $\{X(\omega), \omega \in \Omega\}$, we can make a histogram with bins $0, 1, ..., n$, where bin k collects those ω whose $X(\omega) = k$. We can imagine each ω is a small ball. Then the hight of the pile of balls in the k-th bin is proportional to $\binom{n}{k}/2^n$.

This population histogram is also the probability distribution of the random variable X, which is a random sample from the 2^n numbers $\{X(\omega), \omega \in \Omega\}$.

The center of this distribution is $n/2$. The large deviation result is about the tails of this distribution, i.e., what is the probability that X deviates from its center $n/2$ by at least ϵn?

15 Small deviation

The large deviation result tells us about the probability that X deviates from $n/2$ by at least ϵn. For large n, this probability is extremely small, so the whole bulk of the histogram of $\{X(\omega), \omega \in \Omega\}$ lies within this deviation. We want to zoom in to study smaller deviation.

Let $k = n/2 + z\sqrt{n}/2$, or $z = (k - n/2)/(\sqrt{n}/2)$, let us calculate $P(X = k) = g(z)$. Here the deviation is proportional to $\sqrt{n}$ instead of n.

First, we calculate $p(0) = P(X = n/2)$. Using the fact that $\binom{n}{n\alpha} \sim C(n,\alpha)2^{nh(\alpha)}$ that we derived in Section 8, we have

$$g(0) \sim \frac{1}{\sqrt{2\pi}}\frac{2}{\sqrt{n}}.$$

Next, we calculate $g(z)/g(0)$. Let $d = z\sqrt{n}/2$ be the deviation,

$$\begin{aligned}
\frac{g(z)}{g(0)} &= \binom{n}{n/2+d}/\binom{n}{n/2} \\
&= \frac{(n/2)!(n/2)!}{(n/2+d)!(n/2-d)!} \\
&= \frac{n/2(n/2-1)...(n/2-d+1)}{(n/2+1)(n/2+2)...(n/2+d)} \\
&= \frac{(1-\Delta)(1-2\Delta)...(1-(d-1)\Delta)}{(1+\Delta)(1+2\Delta)...(1+d\Delta)}
\end{aligned}$$

where $\Delta = 2/n$.

16 $1 + \delta \approx e^{\delta}$

According to Taylor expansion, $e^{\delta} = 1 + \delta + O(\delta^2)$, we can simply write $1 + \delta$ as e^{δ} in calculating the limit, thus

$$\begin{aligned}
\frac{g(z)}{g(0)} &\approx \exp\{-\sum_{i=1}^{d-1}(i\Delta) - \sum_{i=1}^{d}(i\Delta)\} \\
&= \exp\{-\Delta[(d(d-1))/2 + (d(d+1))/2]\} = e^{-z^2/2}.
\end{aligned}$$

In the limit, the $\approx$ sign becomes $=$ sign. So for any z, we have

$$g(z) \sim \frac{1}{\sqrt{2\pi}}e^{-z^2/2}\frac{2}{\sqrt{n}}.$$

17 From sum to integral

This approximation can be quite useful for calculating some probabilities. For example, if we flip a fair coin 100 times, what is the probability that the number of heads is between 40 and 60, or what is the probability that the number of heads is greater than 60? In general, suppose we want to compute $P(a \le X \le b)$, where a and b are two integers from 0 and n, we need to calculate the sum

$$P(a \le X \le b) = \sum_{k=a}^{b} P(X = k) = \sum_{k=a}^{b} \binom{n}{k}/2^n,$$

which can be quite difficult to calculate. But if we use the limiting result in Section 16, we can calculate

$$P(a \le X \le b) \sim \sum_{k=a}^{b} g(z) = \sum_{k=a}^{b} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \frac{2}{\sqrt{n}},$$

where $z = (k - n/2)/(\sqrt{n}/2)$. When k runs from a to b, z runs from a' to b', where $a' = (a - n/2)/(\sqrt{n}/2)$ and $b' = (b - n/2)/(\sqrt{n}/2)$, and every two consecutive values of z are apart by $2/\sqrt{n}$. Therefore, as $n \to \infty$,

$$P(a \le X \le b) \to \int_{a'}^{b'} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Let $Z = (X - n/2)/(\sqrt{n}/2)$, that means

$$P(a \le X \le b) = P(a' \le Z \le b') \to \int_{a'}^{b'} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Although Z is discrete, for large n, we may treat Z as a continuous random variable, whose probability density function is $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. This distribution is called a normal distribution or Gaussian distribution.

18 Gaussian distribution

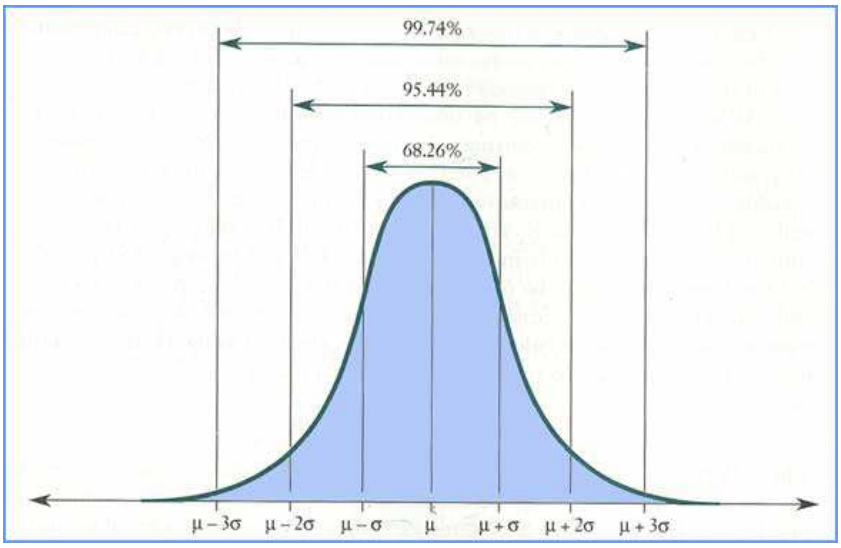

Figure 2: Gaussian distribution with mean μ and variance σ^2.

In general, a Gaussian or normal distribution is denoted by $\mathrm{N}(\mu, \sigma^2)$, where μ is called the mean, and σ^2 is called the variance, and σ is called standard deviation. The probability density function is illustrated by Figure 2. It is a bell shaped curve, centered at μ. The area within 1 standard deviation σ from the center is about 68%. The area within 2σ is about 95%. The area within 3σ is about 99.7%.

The probability density function $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ corresponds to $\mu = 0$ and $\sigma^2 = 1$. It is called standard normal distribution.

Figure 3: Gauss

19 Continuous distribution: sampling from area under curve

A very intuitive way to understand the probability density function $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is through the following thought experiment. We call the area under the curve $f(z)$ the area S. We randomly throw a point into S, and then we return the horizontal coordinate of this random point. This coordinate is a realization of Z. Clearly, the higher the $f(z)$ is, the more likely that Z is around z. Also, from Figure 2, we know that $P(Z \in [-1, 1]) = 68.2\%$, and $P(Z \in [-2, 2]) = 95.4\%$.

20 Continuous distribution: histogram of a population

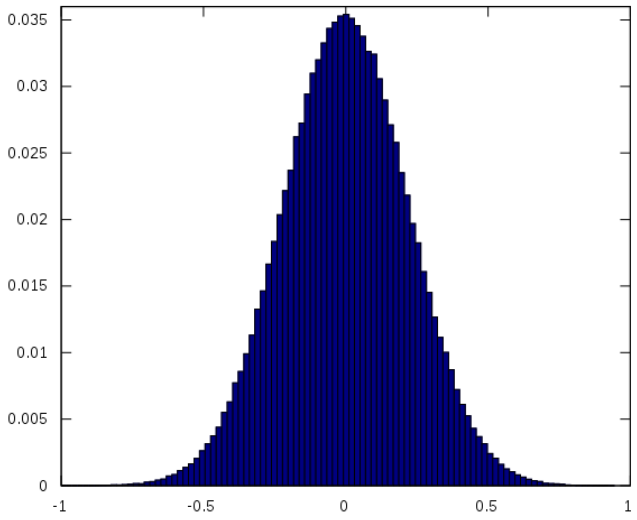

Figure 4: Histogram converges to density.

To understand $f(z)$, we can also consider the population of Z-score, $\{Z(\omega) = (X(\omega) - n/2)/(\sqrt{n}/2), \omega \in \Omega\}$, that is, for each sequence ω, we can calculate its Z-score. So we have 2^n scores. We may consider plotting a histogram of these scores, so that the area of each bin is the proportion of the scores that fall into this bin. If we let $n \to \infty$, and at the same time, let the bin size goes to 0, then the tops of the bins form a continuous curve, and this curve is $f(z)$. See Figure 4 for an illustration.

21 Continuous distribution: population density

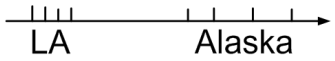

Figure 5: Density of population.

For a small interval around z, say, $[z, z+\Delta z]$, we can count the number of those $Z(\omega)$ that live in this neighborhood $[z, z+\Delta z]$. Let this number be $n(z)$, then the population density is $n(z)/\Delta z$, that is, the number of people divided by the size of the neighborhood. It tells us how densely a local neighborhood is populated. For example, the density of Los Angeles is much greater than the density of Alaska. See Figure 5 for an illustration.

It is better that we normalize the size of the population $N = 2^n$, that is, we use $[n(z)/N]/\Delta$ as the measure of density. As $\Delta z \to 0$, it becomes the density curve $f(z)$.

If we randomly sample Z from the population $\{Z(\omega)\}$, then the population proportion $n(z)/N$ becomes the probability $P(Z \in [z, z+\Delta z])$, and $f(z) = P(Z \in [z, z+\Delta z])/\Delta z$ is the probability density. We use the notation $Z \sim f(z)$ to denote that Z is sampled from $f(z)$.

22 Central limit theorem

The fact that the limiting distribution of $Z = (X - n/2)/(\sqrt{n}/2)$ is a Gaussian distribution is called the central limit theorem.

One may wonder why we have $k = n/2 + z\sqrt{n}/2$ in our study of the central limit theorem? Why $\sqrt{n}/2$? Why not $n^{1/3}$? Why not $\sqrt{n}/3$? This is because $\sqrt{n}/2$ is the average deviation or standard deviation as we will see later.

23 Hypothesis testing

Suppose we flip a coin 100 times, and we get 55 heads. Is the coin fair? Obviously, 55 is not 50. But one can argue that it is rather likely that the number of heads deviate from 50 purely by chance. The question is, whether 55 deviates from 50 too much to be accounted for by chance?

If the hypothesis that the coin is fair is correct, then what do we expect the number of heads? We know that $Z = (X - n/2)/(\sqrt{n}/2) = (X - 50)/5$ should follow a standard normal distribution. Here the observed value of $Z_{\text{obs}} = (55 - 50)/5 = 1$. If we randomly sample Z from $\mathrm{N}(0, 1)$, even by chance, the probability that we get a value that is greater than 1 or less than -1 is about 32%. So Z_{obs} is not extraordinary or suspicious. Therefore, we have no reason to doubt that the coin is fair.

24 Expectation

For the population $\{X(\omega), \omega \in \Omega\}$. If we randomly sample an ω from Ω, then $X(\omega)$ is a random number. Let us define

$$\mu = \mathrm{E}[X] = \langle X \rangle = \frac{1}{N} \sum_{\omega=1}^{N} X(\omega),$$

where $N = 2^n$ is the size of the population, and we label the sequences ω from 1 to N.

We can interpret the expectation as the population average, or as the center of the fluctuation of the random variable X. $\mathrm{E}[X]$ is also the center of the histogram of $\{X(\omega), \omega \in \Omega\}$.

25 Variance

Let us also define

$$\sigma^2 = \mathrm{Var}[X] = \mathrm{E}[(X-\mu)^2] = \frac{1}{N}\sum_{\omega=1}^{N}(X(\omega)-\mu)^2,$$

which is another population average. But it is not the average of $X(\omega)$, but the average of the squared deviation $(X(\omega)-\mu)^2$. We call this average the variance. It can be interpreted as the population variance, or the magnitude of the fluctuation of the random variable X. It measures the spread of the histogram of $\{X(\omega), \omega \in \Omega\}$.

Note that the variance is defined as average squared deviation. So its unit is the square of the unit of X. The square root of the variance, σ, is in the same unit as X, and is called standard deviation.

26 Under linear transformation

If we let $Y(\omega) = aX(\omega) + b$, then

$$\mu_Y = \mathrm{E}[Y] = \frac{1}{N}\sum_{\omega=1}^{N}[aX(\omega)+b] = a\mathrm{E}[X] + b = a\mu_X + b,$$

$$\mathrm{Var}[Y] = \mathrm{E}[(Y-\mu_Y)^2] = \frac{1}{N}\sum_{\omega=1}^{N}[(aX(\omega)+b) - (a\mu_X+b)]^2 = a^2\mathrm{Var}[X].$$

27 Expectation and variance in coin flipping

We shall show that $\mathrm{E}[X] = n/2$ and $\mathrm{Var}[X] = n/4$, then the standard deviation is $\sigma = \sqrt{n}/2$, which is the average deviation. That is why we let $k = n/2 + z\sqrt{n}/2$, that is, we measure the deviation of k from $n/2$ in terms of the average deviation $\sqrt{n}/2$.

The results that $\mathrm{E}[X] = n/2$ and $\mathrm{Var}[X] = n/4$ follow from the recursive relationship: $\mu_{n+1} = \mu_n + 1/2$, and $\sigma^2_{n+1} = \sigma^2_n + 1/4$, where μ_n and σ^2_n are the expectation and variance of X in n flips.

The proof of the recursive relationship is simple. For each ω of n flips, the next flip can either be 1 or 0. That is, each ω spawns two sequences $(\omega, 0)$ and $(\omega, 1)$. So each $X(\omega)$ spawns two numbers $X(\omega)$ and $X(\omega)+1$. Therefore,

$$\begin{aligned}\mu_{n+1} &= \frac{1}{2N}\sum_{\omega=1}^{N}[X(\omega) + (X(\omega)+1)] \\ &= \frac{1}{N}\sum_{\omega=1}^{N}X(\omega) + 1/2 = \mu_n + 1/2.\end{aligned}$$

$$\begin{aligned}\sigma^2_{n+1} &= \frac{1}{2N}\sum_{\omega=1}^{N}[(X(\omega) - (\mu_n+1/2)]^2 + [(X(\omega)+1) - (\mu_n+1/2)]^2 \\ &= \frac{1}{N}\sum_{\omega=1}^{N}[X(\omega)-\mu_n]^2 + 1/4 = \sigma^2_n + 1/4.\end{aligned}$$

28 Law of large number reaffirmed

According to the results in Section 26, $\mathrm{E}[X/n] = \mathrm{E}[X]/n = 1/2$, and $\mathrm{Var}[X/n] = \mathrm{Var}[X]/n^2 = 1/4n \to 0$. That is, the frequency of heads X/n fluctuates around $1/2$, but the magnitude of fluctuation diminishes as $n \to 0$. This reaffirms the weak law of large number.

29 Galton's Quincunx

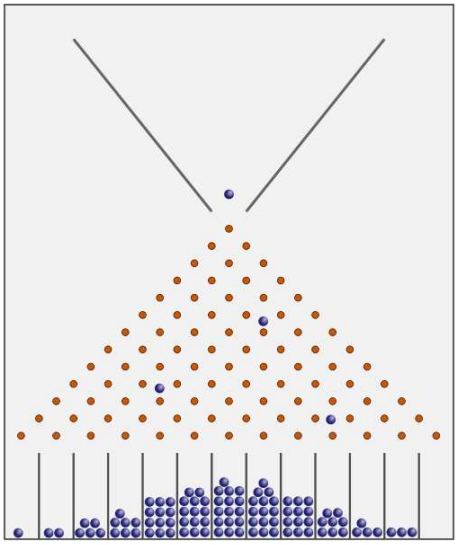

Figure 6: Quincunx.

Figure 6 displays a device called quincunx. The small balls go through such a device and end up in one of the bins at the bottom. At each step, the small ball makes either a left turn or a right turn. Suppose the ball goes through n steps. If we number the locations at which the ball exists the quincunx as 0, 1, 2, ..., n, from left to right. Let X be the exit location of a ball. Then X is the number of heads if we flip a fair coin n times. So this device is essentially the same as coin flipping. For each exit location k, we count the number of paths that end up in k, and this number is $\binom{n}{k}$. We can also calculate this number using Pascal triangle, where each number is the sum of the two numbers on its shoulder.

If we drop a large number of balls into these bins through the quincunx, then according to the central limit theorem, the balls in these bins are in the shape of the normal distribution.

30 A population on the move

Now consider a random walk on integers. At each step, the random walker either goes left or right with $1/2$ probability. This is essentially the same as the quincunx. The only difference is how we label the final location. Let $Z_i = 1$ if the i-th coin flip is head, and $Z_i = -1$ otherwise. Let $X(n) = Z_1 + ... + Z_n$. Then $X(n)$ is the destination of the random walk after n steps. Let $\tilde{X}(n)$ be the number of heads after n flips of the fair coin. Then $X(n) = \tilde{X}(n) - (n - \tilde{X}(n)) = 2\tilde{X}(n) - n$. According to Section 26, $\mathrm{E}[X(n)] = 2\mathrm{E}[\tilde{X}(n)] - n = 0$ and $\mathrm{Var}[X(n)] = 4\mathrm{Var}[\tilde{X}(n)] = n$.

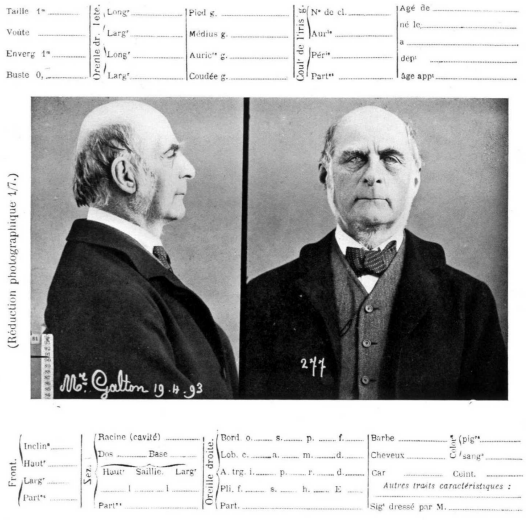

Figure 7: Galton

We can imagine that each integer is a state. Imagine that a population of 1 million people start from state 0. At each step, within each state, half of the people move to the left neighbor, and the other half move to the right neighbor. Then we can imagine that this population gradually diffuses from state 0 to other states.

31 Random walk on graph

Instead of a random walk on all the integers, we may also consider a random walk on a graph.

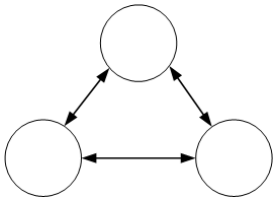

Figure 8: Random walk on three states.

For example, consider a simple graph of three nodes or states, 1, 2, and 3, as illustrated by Figure 8. At each step, the person randomly flips a coin, and goes to one of the other two states with equal probabilities. Let X_t be the state of this person at time $t = 0, 1, ...$, and let us assume that $X_0 = 1$, i.e., the random walker starts from state 1.

32 Transition probability

The transition probability is defined as $K_{ij} = P(X_{t+1} = j | X_t = i)$, where $i, j \in \{1, 2, 3\}$. K_{ij} is conditional probability: given that the person is currently at state i, what is the probability that

he will be in state j at the next time point. In this example, the conditional probability is given, and can be arranged into a matrix:

$i\backslash j$	1	2	3
1	0	1/2	1/2
2	1/2	0	1/2
3	1/2	1/2	0

33 Markov property

Such a random walk satisfies the Markov property, $P(X_{t+1} = j|X_t = i, X_{t-1}, ..., X_0) = P(X_{t+1} = j|X_t = i)$, That is, given the present state $X_t = i$, the future X_{t+1} has nothing to do with the past $X_{t-1}, ..., X_0$. The sequence X_0, X_1, ... forms a Markov chain.

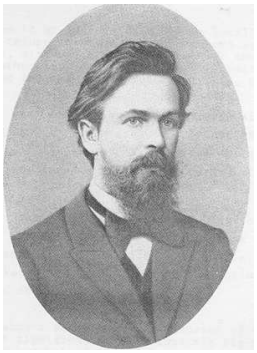

Figure 9: Markov

34 Stationary distribution

Using the formula $P(X_{t+1} = j) = \sum_{i=1}^{3} P(X_t = i)P(X_{t+1} = j|X_t = i)$ recursively, we get the distributions of X_t:

i	1	2	3
$P(X_1 = i)$	0	1/2	1/2
$P(X_2 = i)$	1/2	1/4	1/4
$P(X_3 = i)$	1/4	3/8	3/8
$P(X_4 = i)$	3/8	5/16	5/16
$P(X_5 = i)$	5/16	11/32	11/32

This distribution converges to the uniform distribution over the three states.

35 Population view

Suppose a population of 1 million people start from state 1. At each step, within each state, half of the people goes to one neighbor, and the other half goes to the other neighbor. Then the distributions calculated in Section 34 tells us the distribution of the population at time $t = 1, 2, ...$

The formula $P(X_{t+1} = j) = \sum_{i=1}^{3} P(X_t = i)P(X_{t+1} = j|X_t = i)$ can be interpreted as follows. $P(X_t = i)$ is the number of people (in the unit of million) in state i at time t. $P(X_{t+1} = j|X_t = i)$ is the fraction of those people in state i who will move to state j. $P(X_{t+1} = j)$ is the number of people in state j at time $t + 1$, which is the sum of the people from the three states.

36 Arrow of time

We can imagine that a population of 1 million people start from state 1, and this population gradually diffuses to the three states and eventually uniformly distributed over the three states. After that, the distribution will remains stationary, even though people keep moving around.

Here the movement of each person is time reversible, that is, $P(X_{t+1} = j|X_t = i) = P(X_{t+1} = i|X_t = j)$. However, the population distribution is not time reversible: it converges to a uniform distribution. This is the so-called arrow of time, and it is of a probabilistic nature.

37 Back tracing

We can also calculate the back tracing probability, e.g., $P(X_3 = i|X_4 = j)$ for all pairs of (i, j), using the formula $P(X_3 = i|X_4 = j) = P(X_3 = i, X_4 = j)/P(X_4 = j) = P(X_3 = i)P(X_4 = j|X_3 = i)/P(X_4 = j)$.

$i\backslash j$	1	2	3
1	$\frac{1/4\times 0}{3/8} = 0$	$\frac{1/4\times 1/2}{5/16} = 2/5$	$\frac{1/4\times 1/2}{5/16} = 2/5$
2	$\frac{3/8\times 1/2}{3/8} = 1/2$	$\frac{3/8\times 0}{5/16} = 0$	$\frac{3/8\times 1/2}{5/16} = 3/5$
3	$\frac{3/8\times 1/2}{3/8} = 1/2$	$\frac{3/8\times 1/2}{5/16} = 3/5$	$\frac{3/8\times 0}{5/16} = 0$

The population interpretation is as follows: Suppose a population of 1 million people start from state 1. Among all the people who end up in state j at time 4, what is the fraction of those who were at state i at time 3? $P(X_4 = j)$ can be interpreted the number of people in state j at step 4. $P(X_3 = i, X_4 = j)$ can be interpreted as the number of people who came from state i.

The back tracing probability plays an important role in establishing the arrow of time and in quantifying the speed of convergence of the Markov chain to the stationary distribution.

38 Brownian motion

If we put a drop of milk into a cup of coffee, then the drop of milk diffuses. In the idealized and simplified one-dimensional situation, so that there is no gravity and no interactions between the particles of the milk, we can imagine that each particle of the drop of milk follows a random walk caused by the bombardments of the tiny water molecules. At each moment, the density of the drop of milk is described by the normal distribution. Imagine that we can videotape this process, so that we have a movie.

The motion of a particle caused by the bombardment of the tiny water molecules is called Brownian motion. It was first discovered by Botanist Brown and was later analyzed by Einstein.

39 Making a movie

We can also model this process, as if we are making a movie by simulating from the model instead of videotaping the actual process.

Figure 10: Einstein

We all know that a movie is made up of a finite number of frames, even though we perceive an apparent continuous motion. We can also build our model this way. We divide the time domain into small periods: $(0, \Delta t)$, $(\Delta t, 2\Delta t)$, ... Within each period, the particle moves either to the right or to the left by Δx.

At time t, we have gone through $n = t/\Delta t$ flips. Let $X(t)$ be the location of a particle at time t. According to the calculation in Section 30, $\mathrm{E}[X(t)] = 0$, and $\mathrm{Var}[X(t)] = n\Delta x^2 = t\Delta x^2/\Delta t$.

If we want our model to give a definitive answer as $\Delta t \to 0$, then we need to have $\Delta x^2/\Delta t$ goes to a constant that is independent of Δt or Δx. Let D be this limit, then for small Δt, we have $\Delta x^2 = D\Delta t$, or $\Delta x = \sqrt{D\Delta t}$.

From the point of view of making a movie, we want to see the same scene, i.e., the distribution of the particles, at any fixed time t no matter how many frames we show in a second, or no matter what Δt is. Then at least we want $\mathrm{Var}[X(t)]$ to be the same for different Δt. Therefore, we want $\Delta x^2/\Delta t$ to be a constant.

According to the central limit theorem, $X(t)$ follows a normal distribution with expectation 0 and variance Dt. The constant D is the diffusion constant, which depends on the size of the water molecules and some other physical constants.

40 Velocity not defined

Suppose you want to make a movie of a particle moving at velocity v, then you want to have $\Delta x = v\Delta t$. But in Brownian motion, we have $\Delta x^2 = D\Delta t$, or $\Delta x = \sqrt{D}\sqrt{\Delta t}$, so $\Delta x/\Delta t = \sqrt{D}/\sqrt{\Delta t} \to \infty$ as $\Delta t \to 0$. That is, the velocity of the particle is not defined in Brownian motion, and we see a zigzag path.

The relationship $\Delta x^2 = D\Delta t$ plays a fundamental role in stochastic differential equations.

41 Flip a biased coin

Now we are ready to consider the more general situation where the probability of getting a head is p, which is not necessarily $1/2$. For a sequence $\omega \in \Omega$, we still have $X(\omega) = Z_1(\omega) + ... + Z_n(\omega)$. We use the notation $Z_i \sim \mathrm{Bernoulli}(p)$ to denote the fact that Z_i is a binary random variable, and $P(Z_i = 1) = p = 1 - P(Z_i = 0)$. We write $X \sim \mathrm{Binomial}(n, p)$.

We can show that $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$. The reason is that for each sequence ω with exactly k heads, the probability of this sequence, $P(\omega) = p^k(1-p)^{n-k}$. So this time, the

probability is not uniform. The number of sequences with k heads is $\binom{n}{k}$. So the total probability is $\binom{n}{k}p^k(1-p)^{n-k}$.

42 What is independence

The reason that $P(\omega) = p^k(1-p)^{n-k}$ is because we assume that the n flips are independent, so that we can multiply their probabilities.

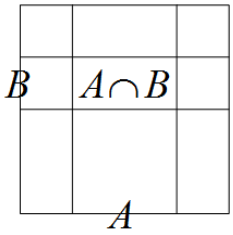

Figure 11: A and B are independent.

Figure 11 illustrates the concept of independence. Suppose we generate two uniform random variables (X, Y) from $[0,\ 1]$ independently. Then (X, Y) is a random point in the unit square $[0, 1]^2$. Let $A \subset [0, 1]$ be an interval on the x-axis, and $B \subset [0, 1]$ be an interval on the y-axis. So $P(A) = P(X \in A) = |A|$, and $P(B) = P(Y \in B) = |B|$, where $|A|$ and $|B|$ are the lengths of A and B. With a little abuse of notation, we can also think of A and B as the corresponding vertical and horizontal strips respectively. Then $A \cap B$ is the rectangular intersection between the vertical strip A and the horizontal strip B. $P(A \cap B) = P((X, Y) \in A \cap B) = |A \cap B| = |A||B| = P(A)P(B)$, where $|A \cap B|$ is the area of $A \cap B$.

So the geometrical meaning that two events are independent is that they are perpendicular to each other.

43 Map to equal likely setting: survey sampling

In the experiment of flipping a biased coin, the outcomes are not equally likely, because if a sequence ω has k heads, then $P(\omega) = p^k(1-p)^{n-k}$.

But we can map this experiment to another experiment where the outcomes are still equally likely. Consider sampling from a population of M people, where r of them are red, and b of them are blue, and $r + b = M$. If we randomly sample a person, then the probability we get a red person is $p = r/M$.

If we sequentially sample n people independently and with replacement, that is, after we sample a person, we put the person back to the population, so that the person can still be sampled later on. In this scenario, the sample space Ω consists of all the sequences of people that we sampled. The number of all the possible sequences is M^n, because each time, we have M different choices. All these sequences are equally likely.

The number of sequences with k red people is $\binom{n}{k}r^k b^{n-k}$. The reason is as follows. In the sequence of n people, we can choose k of them to be red, and then the rest of $n - k$ of them to be blue. The number of choices is $\binom{n}{k}$. Then for each red person, we have r choices in random sampling, and for each blue person, we have b choices.

Let B_k be the subset of all the sequences with k red people. Then $P(B_k) = |B_k|/|\Omega| = \binom{n}{k} r^k b^{n-k}/M^n = \binom{n}{k} p^k (1-p)^{n-k}$. This is the same formula as we derived for coin flipping experiment.

44 How large is $P(A_k)$?

Now let us go back to the coin flipping experiment. Again let $A_k = \{\omega : X(\omega) = k\}$, i.e., the set of sequences with k heads. According to Section 41, $P(A_k) = \binom{n}{k} p^k (1-p)^{n-k}$.

In Section 8, we have shown that for $k = n\alpha$,

$$\binom{n}{k} \sim C(n,\alpha) 2^{nh(\alpha)},$$

where

$$C(n,\alpha) = \frac{1}{\sqrt{2\pi n\alpha(1-\alpha)}},$$
$$h(\alpha) = -[\alpha \log_2 \alpha + (1-\alpha)\log_2(1-\alpha)].$$

So

$$\begin{aligned} P(A_k) = \binom{n}{k} p^k (1-p)^{n-k} &\sim C(n,\alpha) 2^{nh(\alpha)} p^{n\alpha} (1-p)^{n(1-\alpha)} \\ &= C(n,\alpha) 2^{nh(\alpha)} 2^{n(\alpha \log_2 p + (1-\alpha)\log_2(1-p))} \\ &= C(n,\alpha) 2^{-nD(\alpha)} \end{aligned}$$

where

$$D(\alpha) = [\alpha \log_2 \frac{\alpha}{p} + (1-\alpha)\log_2 \frac{1-\alpha}{1-p}].$$

45 Kullback-Leibler divergence

$D(\alpha)$ can be written as $D((\alpha, 1-\alpha)||(p, 1-p))$, which is the Kullback-Leibler divergence of the probability distribution $(\alpha, 1-\alpha)$ from the probability distribution $(p, 1-p)$.

As a function of α, $D(\alpha) \geq 0$, and $D(\alpha) = 0$ only when $\alpha = p$. This is because $D'(\alpha) = \log_2(\alpha/p) - \log_2((1-\alpha)/(1-p))$, and $D''(\alpha) = 1/\alpha + 1/(1-\alpha)$. For $\alpha \in (0,1)$, $D'' > 0$, so the function is convex, and the minimum is achieved as $\alpha = p$, where $D'(\alpha) = 0$, and $D(\alpha) = 0$.

46 Concentration of probability

Because $P(A_k) \sim C(n,\alpha) 2^{-nD(\alpha)}$, and $D(\alpha)$ achieves its minimum 0 at $\alpha = p$, the probability will eventually concentrate on those classes A_k where $k/n = \alpha \approx p$. Because if $k/n = \alpha$ is deviate from p by a constant $\epsilon > 0$, then $P(A_k) \sim C(n,\alpha) 2^{-nD(\alpha)}$ goes to 0 at an exponential rate no matter how small ϵ is.

47 Weak law of large number

Because $P(A_k) = P(X = k)$, we have the following weak law of large number

$$P(|\frac{X}{n} - p| > \epsilon) \to 0.$$

48 Probability and long run frequency

The weak law of large number says that the long run frequency converges to probability. In other words, we can interpret probability as long run frequency.

But bear in mind that we never define probability as long run frequency. In fact, such a definition is untenable. If we define $p = \lim_{n\to\infty} X/n$, we will face the question whether this limit exists at all. In fact, it is possible for X/n to be as small as 0 or as large as 1. You may argue that the probability for X/n to be 0 or 1 is extremely small. But then you have not define probability yet, how can you talk about probability being extremely small? As a matter of fact, the statement that $X/n \to p$ has to be expressed in terms of probability, as in the weal law of large number.

So the fact that probability manifests itself as long run frequency is a logical consequence, not a logical starting point.

In the equally likely setting, we can define probability $P(A) = |A|/|\Omega|$. For instance, in the survey sampling situation, all the M^n sequences are equally likely. The weak law of large number corresponds to the following mathematical fact: among all these M^n sequences, if we look at the frequency of red people in each sequence, then almost all of M^n sequences produce frequencies that are close to $p = r/M$.

49 Mean and variance of Binomial

For each sequence ω, we can calculate $X(\omega)$, which is the number of heads. If $X(\omega) = k$, the probability of this sequence is $P(\omega) = p^k(1-p)^k$.

We can define the expectation

$$\mu = \mathrm{E}[X] = \sum_{\omega=1}^{N} X(\omega)P(\omega),$$

where again we number the sequences from 1 to $N = 2^n$. This time, $\mathrm{E}[X]$ is a weighted average, where the weight of ω is its probability.

We can also define the variance

$$\sigma^2 = \mathrm{Var}[X] = \sum_{\omega=1}^{N} (X(\omega) - \mu)^2 P(\omega).$$

Variance is the average of the squared deviation.

Because $X = Z_1 + ... + Z_n$, where $Z_i \sim \text{Bernoulli}(p)$ independently, we have

$$\begin{aligned} \mathrm{E}[X] &= \sum_{i=1}^{n} \mathrm{E}[Z_i] = np, \\ \mathrm{Var}[X] &= \sum_{i=1}^{n} \mathrm{Var}[Z_i] = np(1-p). \end{aligned}$$

The above results follow from the following two facts. The first fact is that for a random variable $Z \sim \text{Bernoulli}(p)$, $\mathrm{E}[Z] = 0 \times (1-p) + 1 \times p = p$, and $\mathrm{Var}[Z] = (0-p)^2 \times (1-p) + (1-p)^2 \times p = p(1-p)$. The second fact is that when we calculate the expectation and variance of the sum of some independent random variables, we can simply add up the expectations and variances of these random variables.

50 Sum of two random variables

Suppose we have two sample spaces, e.g., two boxes. In one box, we have N black ball, and we number them $1, 2, ..., N$. Each black ball carries a number, and let the number of black ball i be $X(i)$. Suppose when we sample from this box, the probability of getting the black ball i is $p(i)$. In the other box, we have M white balls, and we number them $1, 2, ..., M$. Each white ball carried a number too, and let the number of white ball j be $Y(j)$. Suppose when we sample from this box, the probability of getting the white ball j is $q(j)$.

Suppose we sample a black ball from the first box, and independently, we sample a white ball from the second box. Then we sum the number carried by the black ball and the number carried by the white ball. Let the sum be Z. We want to study the behavior of this random variable Z.

The joint sample space is $\Omega = \{1, 2, ..., N\} \times \{1, 2, ..., M\}$, i.e., a $N \times M$ table of pairs. For each pair $\omega = (i, j) \in \Omega$, we get two numbers $X(i)$ and $Y(j)$, and their sum is $Z(\omega) = X(i) + Y(j)$. Because of the independence, the probability of getting a pair $\omega = (i, j)$ is $P(\omega) = p(i)q(j)$.

$$\begin{aligned}
\mu_Z = \mathrm{E}[Z] &= \sum_{\omega\in\Omega} Z(\omega)P(\omega) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}[X(i) + Y(j)]p(i)q(j) \\
&= \sum_{i=1}^{N} X(i)p(i) + \sum_{j=1}^{M} Y(j)q(j) \\
&= \mathrm{E}[X] + \mathrm{E}[Y] = \mu_X + \mu_Y.
\end{aligned}$$

$$\begin{aligned}
\sigma_Z^2 = \mathrm{Var}[Z] &= \sum_{\omega\in\Omega} (Z(\omega) - \mu_Z)^2 P(\omega) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}[(X(i) + Y(j)) - (\mu_X + \mu_Y)]^2 p(i)q(j) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}[(X(i) - \mu_X)^2 + (Y(j) - \mu_Y)^2 + 2(X(i) - \mu_X)(Y(j) - \mu_Y)]p(i)q(j) \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y].
\end{aligned}$$

51 Law of large number

According to the results in Section 49, $\mathrm{E}[X/n] = \mathrm{E}[X]/n = p$ and $\mathrm{Var}[X/n] = \mathrm{Var}[X]/n^2 = p(1-p)/n \to 0$ as $n \to \infty$. Therefore, $X/n \to p$. This reaffirms the law of large number.

52 Central limit theorem

Let $Z = (X - np)/\sqrt{np(1-p)}$, then using the results in Section 26, $\mathrm{E}[Z] = (\mathrm{E}[X] - np)/\sqrt{np(1-p)} = 0$ and $\mathrm{Var}[Z] = \mathrm{Var}[X]/(np(1-p)) = 1$. We can show that the distribution of Z goes to a Gaussian distribution, using the same method as in Sections 15 and 16.

For $k = np + z\sqrt{np(1-p)}$, or $z = (k-np)/\sqrt{np(1-p)}$, and let $g(z) = P(X = k)$. Then according to the result in Section 44

$$g(0) = P(X = np) \sim C(n,p)2^{-nD(p)} = \frac{1}{\sqrt{2\pi np(1-p)}},$$

because $D(p) = 0$. Let $d = z\sqrt{np(1-p)}$, and let $q = 1-p$,

$$\begin{aligned}
\frac{g(z)}{g(0)} &= \left(\binom{n}{np+d}p^d\right) / \left(\binom{n}{np}q^d\right) \\
&= \frac{(np)!(nq)!p^d}{(np+d)!(nq-d)!q^d} \\
&= \frac{nq(nq-1)...(nq-d+1)p^d}{(np+1)(np+2)...(np+d)q^d} \\
&= \frac{(1-\Delta_0)(1-2\Delta_0)...(1-(d-1)\Delta_0)}{(1+\Delta_1)(1+2\Delta_1)...(1+d\Delta_1)}
\end{aligned}$$

where $\Delta_0 = 1/(nq)$ and $\Delta_1 = 1/(np)$. Using $1+\delta \approx e^\delta$,

$$\begin{aligned}
\frac{g(z)}{g(0)} &\approx \exp\{-\sum_{i=1}^{d-1}(i\Delta_0) - \sum_{i=1}^{d}(i\Delta_1)\} \\
&\approx \exp\{-\Delta_0 d^2/2 - \Delta_1 d^2/2\} = e^{-z^2/2}.
\end{aligned}$$

So for any z, we have

$$p(z) \sim \frac{1}{\sqrt{2\pi np(1-p)}}e^{-z^2/2} = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}\Delta z,$$

where $\Delta z = 1/\sqrt{np(1-p)}$, which is the space between two adjacent values that z can take. Therefore, using the same argument as in Section 17, the random variable $Z = (X-np)/\sqrt{np(1-p)}$ follows the standard normal distribution $\mathrm{N}(0,1)$.

The result is due to de Moivre, who worked as a consultant for gamblers.

Figure 12: de Moivre

53 Survey sampling: margin of error

According to Section 43, suppose we randomly sample n people from a large population of M people, where r of them are red, and b of them are blue. Let X be the number of red people we get. Then $X \sim \text{Binomial}(n, p = r/M)$. If n is large, then $Z = (X - np)/\sqrt{np(1-p)} \sim \text{N}(0,1)$, and $P(Z \in [-2, 2]) = 95\%$. The statement $Z \in [-2, 2]$ is equivalent to $X/n \in [p - 2\sqrt{p(1-p)/n}, p + 2\sqrt{p(1-p)/n}]$, so

$$P(X/n \in [p - 2\sqrt{p(1-p)/n}, p + 2\sqrt{p(1-p)/n}]) = 95\%.$$

If we use $\hat{p} = X/n$ as an estimate of p, which is unknown but is of interest, then with 95% chance, $\hat{p}$ is within $2\sqrt{p(1-p)/n}$ from p. We call $2\sqrt{p(1-p)/n}$ the margin of error, and it can be estimated by $2\sqrt{\hat{p}(1-\hat{p})/n}$.

So 95% times, the estimate $\hat{p}$ is within the margin of error from the true value of p. In other words, 1 out of 20 times, the polls that you see in the newspapers can be wrong.

The interval $[\hat{p} - 2\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + 2\sqrt{\hat{p}(1-\hat{p})/n}]$ is called 95% confidence interval, because it covers the true value p 95% of times.

54 Geometric waiting time

Let T be the number of flips until we get the first head, then $T \sim \text{Geometric}(p)$, $P(T = k) = (1-p)^{k-1}p$, where $k = 1, 2, ...$ This is because the statement $T = k$ is equivalent to the statement that the first $k - 1$ flips are tails and the k-th flip is head.

Let $q = 1 - p$, then

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

55 Poisson process

Suppose we divide the time axis into small periods $(0, \Delta t)$, $(\Delta t, 2\Delta t)$, ... Within each period, we flip a coin independently. Suppose the probability of getting a head is p.

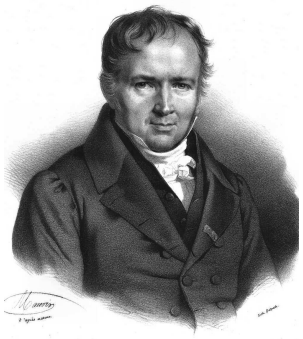

Figure 13: Poisson

Suppose t is a multiple of Δt. Let X be the number of heads within the interval $[0, t]$. Then $X \sim \text{Binomial}(n = t/\Delta t, p)$. $\text{E}[X] = np = tp/\Delta t$. For the model to make sense, we want $p/\Delta t$ goes to a well defined limit as $\Delta t \to 0$. Let λ be this limit, then $p = \lambda \Delta t$, and $\text{E}[X] = \lambda t$. So $\lambda = \text{E}[X]/t$, and it can be interpreted as the number of occurrences (in this case, heads) per unit time.

For $k = 0, 1, ...,$

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

as $\Delta t \to 0$. In the above derivation, we use the result that $(1 - \lambda \Delta t)^{t/\Delta t} \approx (e^{-\lambda \Delta t})^{t/\Delta t} = e^{-\lambda t}$.

56 Exponential waiting time

Let T be the waiting time until the first head, then

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

so

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

This gives us the probability density function of T, as illustrated by Figure 14. Again, we can imagine that T is the horizontal coordinate of a random point under the curve.

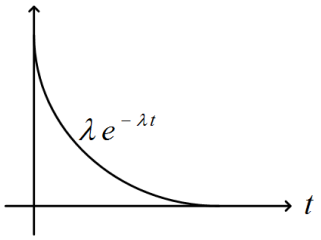

Figure 14: Exponential density function.

The survival probability is

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

This is because $T > t$ means that all the flips up to t give us tails. $P(T > t)$ is the tail area under the density curve in Figure 14.

The expected number of flips until the first head is $1/p = 1/(\lambda\Delta t)$. Each flip takes a duration of Δt. So $\mathrm{E}[T] = 1/(\lambda\Delta t) \times \Delta t = 1/\lambda$.

We can model the time until a particle decays by the exponential model. The half life t is such that $P(T > t) = 1/2$, i.e., half of the particles would decay by the time t.

57 Non-constant intensity and survival function

Suppose the probability of getting a head in $(s, s + \Delta t)$ is $\lambda(s)\Delta t$, which does not have to be constant. Then

$$P(T > t) = \prod_{s=0}^{t} (1 - \lambda(s)\Delta t) \approx \prod_{s=0}^{t} e^{-\lambda(s)\Delta t} \to e^{-\int_0^t \lambda(s)ds}.$$

$S(t) = P(T > t)$ is called survival function. This model is very useful for the so-called survival analysis, i.e., the analysis of the survival times of patients under certain medical treatments.

58 Randomness and entropy

Consider the following distribution on an alphabet of four letters,

x	A	B	C	D
$p(x)$	1/2	1/4	1/8	1/8

We can generate a letter from the above distribution by flipping a fair coin. The scheme is as follows. We flip the coin, if it is head, then we return A. If it is tail, we flip the coin again. If it is head, we return B. If it is tail, we flip the coin a third time. If it is head, we return C. Otherwise, we return D. Figure 15 illustrates this scheme.

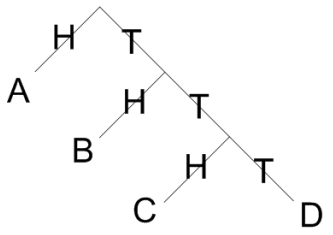

Figure 15: Simulating a random letter by coin flipping.

The number of coin flips for each letter is

x	A	B	C	D
$l(x)$	1	2	3	3

So the average number of coin flips is $1\times1/2+2\times1/4+3\times1/8+3\times1/8 = 7/4$. This is a measure of randomness in the distribution $p(x)$. In general, we define $\text{entropy}(p) = \sum_x [-\log_2 p(x)]p(x)$ as the randomness in a distribution $p(x)$. It is also called Shannon entropy.

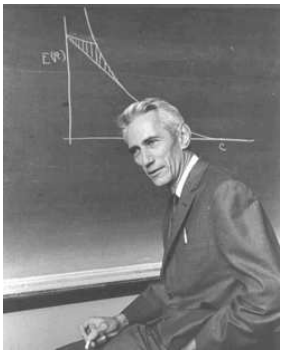

Figure 16: Shannon

59 Coding

What is interesting is that the above scheme leads to a code book for the alphabet:

x	A	B	C	D
code	1	01	001	000

We use 1 for head and 0 for tail. The code book has the property that the code of any letter is not the beginning part of the code of any other letter. So if we have a sequence of letters such as AABDCA, we can code it by 11010000011, from which we can get back the original sequence without ambiguity. The average coding length is just the average number of coin flips, 7/4. So the entropy can be interpreted as the average coding length. It is measured in bits.

60 Coding sequence and random sequence

If we generate a sequence of letters according to $p(x)$ in Section 58, and we code the sequence of letters using the code book in Section 59, then we have a binary sequence, which is also a sequence obtained by flipping a fair coin independently, because the coding scheme in Section 59 comes from the coin flipping scheme in Section 58 for generating the letter.

61 Random sequence is not compressible

In general, a long sequence of random coin flipping is not compressible. We can argue this point by contradiction.

A random sequence is a random draw from the population of 2^n sequences. Suppose a tiny proportion of it is compressible, say, 1% of it can be compressed by less than ρn bits, where $\rho < 1$. Then $1\% \times 2^n \leq 2^{\rho n}$, where the right hand side is the number of sequences with ρn bits, i.e., we can code at most $2^{\rho n}$ different elements with less than ρn bits. But for large n, this is impossible because the ratio between the left hand side and the right hand side goes to infinity.

Therefore, the coding scheme in Section 59 is optimal, in the sense that for a long sequence of letters generated by the distribution $p(x)$, the corresponding binary code is the shortest, in the sense that it cannot be further compressed.

62 Kolmogorov complexity

We can understand randomness from the perspective of coding. We may ask ourselves: why a long sequence 111...1, i.e., the sequence of n 1s, is not random? One argument is that this sequence can be produced by a simple computer program:

```
for i = 1 to n
       write 1
end
```

In computer, the above code can be stored by a short binary sequence. So the original sequence can be compressed into a much shorter sequence.

In general, for any long sequence, we can imagine the shortest code that reproduces this sequence. The length of this shortest code is the Kolmogorov complexity.

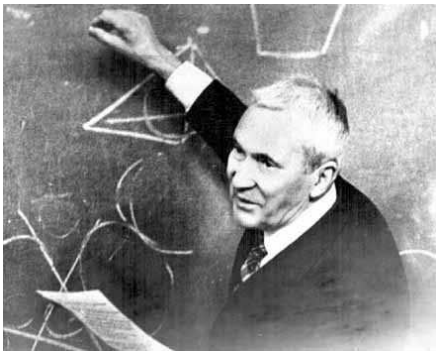

Figure 17: Kolmogorov

63 Randomness = incompressibility

For some long sequences, the shortest codes may simply be the original sequences themselves, that is, the sequences are not compressible. Such sequences are random sequences, because they have all the statistical properties of a random sequence. For instance, we know that a random sequence has its frequency of heads very close to 1/2. If a sequence is not compressible, then its frequency of heads must be close to 1/2.

We can argue this point by contradiction. Suppose the frequency of heads in this sequence is 1/3. Then this sequence belongs to class $A_{n/3}$, and we know that $|A_{n/3}| \sim C(n, 1/3)2^{nh(1/3)}$, where $h(1/3) < 1$. Therefore, we can reproduce this sequence by the following program: "write the i-th sequence in $A_{n/3}$," where in $A_{n/3}$, we can sort the sequences from smallest to largest. Clearly, $i \leq |A_{n/3}|$, and we can code i by $\log_2 |A_{n/3}|$ bits, and $\log_2 |A_{n/3}|/n \to h(1/3) < 1$, so we can code this sequence using less than n bits.

64 Randomness in flipping a biased coin

Suppose we flip a biased coin n times independently, where the probability of head is p. According to the law of large number, the probability concentrates on those classes A_k where k is close to np. The size of $A_{np} \sim C(n, p)2^{nh(p)}$. In order to generate a random member in A_k where k is close to

np, we only need to flip a fair coin $nh(p)$ times. Therefore, the randomness in flipping a biased coin n times amounts to the randomness of flipping a fair coin $nh(p)$ times. So the randomness in flipping a biased coin with probability p is $h(p)$.

65 Equipartition

Another way to think about the above issue is that for a sequence $Z_1, Z_2, ..., Z_n$, where $Z_i = 1$ with probability p and $Z_i = 0$ with probability $1 - p$, the probability of this sequence $Z_1, Z_2, ..., Z_n$ is $p^X(1-p)^{n-X}$, where $X = \sum_i Z_i$. According to the law of large number, $X/n \to p$. Therefore, $P(Z_1, ..., Z_n) \approx p^{np}(1-p)^{n(1-p)} = 2^{-nh(p)}$, which does not depend on $Z_1, ..., Z_n$. That means that $Z_1, ..., Z_n$ is almost like a uniform random sample, whose probability is the same as the probability of a random sequence generated by flipping a fair coin $nh(p)$ times.

The fact that $Z_1, ..., Z_n$ behaves like a uniform random sample is called equipartition property.

The equipartition property was due to Shannon.

66 Strong law of large number

The weak law of large number takes the form $P(|X/n - p| > \epsilon) \to 0$. The statement or the event $|X/n - p| > \epsilon$ only concerns the first n flips, where Ω is the set of all the 2^n sequences. After calculating its probability, we let n goes to infinity.

The strong law is like saying $P(X/n \to p) = 1$. The statement or the event $X/n \to p$ concerns the infinite sequence. In this case, Ω is the set of all the infinite binary sequences, i.e., $\Omega = \{0,1\}^\infty$. So a more precise way of saying $P(X/n \to p) = 1$ is $P(\{\omega : X_n(\omega)/n \to p\}) = 1$, where $X_n(\omega)$ is the number of heads in the first n flips of ω.

The statement $X/n \to p$ means that for any $\epsilon > 0$, there exists an N, so that for any $n \geq N$, $|X_n(\omega)/n - p| \leq \epsilon$. Let $B_n = \{\omega : |X_n(\omega)/n - p| \leq \epsilon\}$ be the set of all sequences such that $|X_n(\omega)/n - p| \leq \epsilon$ is true. Then the statement $X/n \to p$ can be expressed by $\cup_{N=1}^\infty \cap_{n=N}^\infty B_n$, and the strong law means that $P(\cup_{N=1}^\infty \cap_{n=N}^\infty B_n) = 1$.

67 Measure

Although $\cup_{N=1}^\infty \cap_{n=N}^\infty B_n$ is only a subset of Ω, its probability is 1, which is the same as the probability of Ω. This is possible, because the complement of $\cup_{N=1}^\infty \cap_{n=N}^\infty B_n$, although not empty, may have probability 0.

Probability is like a measure, such as length, area, or volume. It measures the size of the subsets or events. It is possible that a non-empty set has measure 0.

For instance, the length of the interval $[0, 1]$ is 1. But the length of the subset $\{.5\}$ is 0, because $\{.5\}$ is just a point. Similarly, the length of all the rational numbers in $[0, 1]$ is also zero, because the set of all the rational numbers is a countable list of points.

The measure-theoretical treatment to probability is due to Kolmogorov.

68 Infinite additivity

For a statement A about the first n flips, we can define its probability as $P(A) = |A|/2^n$, where $|A|$ is the number of sequences that satisfy A, among all the 2^n sequence. However, how is possible for

us to talk about the probability of the statement like $\cup_{N=1}^{\infty} \cap_{n=N}^{\infty} B_n$, which involves the infinite sequences, and which is a logical combination of infinite many statements B_n, $n = 1, 2, ...$?

It is possible to talk about and calculate such a probability because of the so-called infinite additivity. The following are the three axioms of probability:

(1) For any $A \subset \Omega$, $P(A) \geq 0$.

(2) $P(\Omega) = 1$.

(3) For $A_1, A_2, ...$, if $A_i \cap A_j = \phi$, i.e., the empty set, for any $i \neq j$, then $P(\cup_i A_i) = \sum_i P(A_i)$.

The first two axioms are easy to understand. The third one is called infinite additivity, which says that the probability measure of the union of disjoint pieces is the sum of the probability measures of individual pieces. This axiom looks quite natural for measures like length, area, and volume.

Suppose we assume finite additivity first, i.e., $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$. Then infinite additivity assumes something extra: $P(\lim_{n\to\infty} \cup_{i=1}^{n} A_i) = \lim_{n\to\infty} P(\cup_{i=1}^{n} A_i)$. That is, we can smuggle out the limit from inside probability to outside probability. This smuggling enables us to put probability on events like $\cup_{N=1}^{\infty} \cap_{n=N}^{\infty} B_n$ in strong law of large number.

69 σ-algebra and meaningful statements

For a sample space Ω as big as $\{0, 1\}^{\infty}$, we cannot put probability measure on all its subsets, i.e., not all the events are meaningful. In fact, probability is only defined on those meaningful events or statements. The set of all the meaningful statements is called a σ-algebra.

We can construct a σ-algebra as follows. We start from elementary or basic meaningful statements, such as statements about the first n flips, where n is finite. Then we can combine the basic meaningful statements by AND, OR, and NOT, or equivalent, by intersection, union, and complement. For AND and OR, we are allowed to combine infinite but countably many basic statements. For instance, $\cup_{N=1}^{\infty} \cap_{n=N}^{\infty} B_n$ is a meaningful statement constructed from the basic statements B_n. The collection of statements constructed this way does not include all the subsets of the sample space $\Omega = \{0, 1\}^{\infty}$. But this collection is large enough for our purpose.

The infinite additivity enables us to calculate probability of $\cup_{N=1}^{\infty} \cap_{n=N}^{\infty} B_n$ from the probabilities of basic statements B_n, $n = 1, 2, ...$

70 Stigler's law of eponymy

"No scientific discovery is named after its original discoverer." The law may apply to Gaussian distribution, Poisson process, Markov chain, etc. It also applies to Stigler himself, who attributes the discovery of Stigler's Law to Merton.

71 Acknowledgement

Many of the pictures were downloaded from the internet.