Probability Theory

Course Notes — Harvard University — 2011

C. McMullen

May 4, 2011

Contents

These course notes accompany Feller, *An Introduction to Probability Theory and Its Applications*, Wiley, 1950.

I The Sample Space

Some sources and uses of randomness, and philosophical conundrums.

1. Flipped coin.
2. The interrupted game of chance (Fermat).
3. The last roll of the game in backgammon (splitting the stakes at Monte Carlo).
4. Large numbers: elections, gases, lottery.
5. True randomness? Quantum theory.
6. Randomness as a model (in reality only one thing happens). Paradox: what if a coin keeps coming up heads?
7. Statistics: testing a drug. When is an event good evidence rather than a random artifact?
8. Significance: among 1000 coins, if one comes up heads 10 times in a row, is it likely to be a 2-headed coin? Applications to economics, investment and hiring.
9. Randomness as a tool: graph theory; scheduling; internet routing.

We begin with some previews.

Coin flips. What are the chances of 10 heads in a row? The probability is 1/1024, less than 0.1%. Implicit assumptions: no biases and independence. What are the chance of heads 5 out of ten times? ($\binom{10}{5} = 252$, so 252/1024 = 25%).

The birthday problem. What are the chances of 2 people in a crowd having the same birthday? Implicit assumptions: all 365 days are equally likely; no leap days; different birthdays are independent.

Chances that no 2 people among 30 have the same birthday is about 29.3%. Back of the envelope calculation gives $-\log P = (1/365)(1 + 2 + \cdots + 29) \approx 450/365$; and $\exp(-450/365) = 0.291$.

Where are the Sunday babies? US studies show 16% fewer births on Sunday.

Does this make it easier or harder for people to have the same birthday?

The Dean says with surprise, I've only scheduled 20 meetings out-of-town for this year, and already I have a conflict. What faculty is he from?

Birthdays on Jupiter. A Jovian day is 9.925 hours, and a Jovian year is 11.859 earth years. Thus there are $N = 10,467$ possible birthdays on Jupiter. How big does the class need to be to demonstrate the birthday paradox?

It is good to know that $\log(2) = 0.693147\ldots \approx 0.7$. By the back of the envelope calculation, we want the number of people k in class to satisfy $1 + 2 + \ldots + k \approx k^2/2$ with $k^2/(2N) \approx 0.7$, or $k \approx \sqrt{1.4N} \approx 121$.

(Although since Jupiter's axis is only tilted 3°, seasonal variations are much less and the period of one year might have less cultural significance.)

The rule of seven. The fact that $10\log(2)$ is about 7 is related to the 'rule of 7' in banking: to double your money at a (low) annual interest rate of k% takes about $70/k$ years (not $100/k$).

(A third quantity that is useful to know is $\log 10 \approx \pi$.)

The mailman paradox. A mailman delivers n letters at random to n recipients. The probability that the first letter goes to the right person is $1/n$, so the probability that it doesn't is $1 - 1/n$. Thus the probability that no one gets the right letter is $(1 - 1/n)^n \approx 1/e = 37\%$.

Now consider the case $n = 2$. Then he either delivers the letters for A and B in order (A, B) or (B, A). So there is a 50% chance that no one gets the right letter. But according to our formula, there is only a 25% chance that no one gets the right letter.

What is going on?

Outcomes; functions and injective functions. The porter's deliveries are described by the set S of all functions $f : L \to B$ from n letters to n boxes. The mailman's deliveries are described by the space $S' \subset S$ of all $1 - 1$ functions $f : L \to B$. We have $|S| = n^n$ and $|S'| = n!$; they give different statistics for equal weights.

The sample space. The collection S of all possible completely specified outcomes of an experiment or task or process is called the *sample space*.

Examples.

1. For a single toss at a dartboard, S is the unit disk.
2. For the high temperature today, $S = [-50, 200]$.
3. For a toss of a coin, $S = \{H, T\}$.

4. For the roll of a die, $S = \{1, 2, 3, 4, 5, 6\}$.

5. For the roll of two dice, $|S| = 36$.

6. For n rolls of a die, $S = \{(a_1, \ldots, a_n) : 1 \leq a_i \leq 6\}$; we have $|S| = 6^n$. More formally, S is the set of functions on $[1, 2, \ldots, n]$ with values in $[1, 2, \ldots, 6]$.

7. For shuffling cards, $|S| = 52! = 8 \times 10^{67}$.

An *event* is a subset of S. For example: a bull's eye; a comfortable day; heads; an odd number on the die; dice adding up to 7; never getting a 3 in n rolls; the first five cards form a royal flush.

Logical combinations of events correspond to the operators of set theory. For example:

$$A' = \text{not } A = S - A; \quad A \cap B = A \text{ and } B; \quad A \cup B = A \text{ or } B.$$

Probability. We now focus attention on a *discrete sample space*. Then a *probability measure* is a function $p : S \to [0, 1]$ such that $\sum_S p(s) = 1$. Often, out of ignorance or because of symmetry, we have $p(s) = 1/|S|$ (all samples have equal likelihood).

The probability of an *event* is given by

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

If all s have the same probability, then $P(A) = |A|/|S|$.

Proposition I.1 *We have* $P(A') = 1 - P(A)$.

Note that this formula is not based on intuition (although it coincides with it), but is derived from the definitions, since we have $P(A) + P(A') = P(S) = 1$. (We will later treat other logical combinations of events).

Example: mail delivery. The pesky porter throws letters into the pigeonholes of the n students; S is the set of all functions $f : n \to n$. While the mixed-up mailman simple chooses a letter at random for each house; S is the space of all *bijective* functions $f : n \to n$. These give *different answers* for the probability of a successful delivery. (Exercise: who does a better job for $n = 3$? The mailman is more likely than the porter to be a complete failure — that is, to make no correct delivery. The probability of failure for the porter is $(2/3)^3 = 8/27$, while for the mailman it is $1/3 = 2/3! = 8/24$.

This trend continues for all n, although for both the probability of complete failure tends to $1/e$.)

An infinite sample space. Suppose you flip a fair coin until it comes up heads. Then $S = \mathbb{N} \cup \{\infty\}$ is the number of flips it takes. We have $p(1) = 1/2$, $p(2) = 1/4$, and $\sum_{n=1}^{\infty} 1/2^n = 1$, so $p(\infty) = 0$.

The *average* number of flips it takes is $E = \sum_1^\infty n/2^n = 2$. To evaluate this, we note that $f(x) = \sum_1^\infty x^n/2^n = (x/2)/(1 - x/2)$ satisfies $f'(x) = \sum_1^\infty nx^{n-1}/2^n$ and hence $E = f'(1) = 2$. This is the method of *generating functions*.

Benford's law. The event that a number X begins with a 1 in base 10 depends only on $\log(X) \bmod 1 \in [0, 1]$. It occurs when $\log(X) \in [0, \log 2] = [0, 0.30103\ldots]$. In some settings (e.g. populations of cities, values of the Dow) this means there is a 30% chance the first digit is one (and only a 5% chance that the first digit is 9).

For example, once the Dow has reached 1,000, it must double in value to change the first digit; but when it reaches 9,000, it need only increase about 10% to get back to one.

II Elements of Combinatorial Analysis

Basic counting functions.

1. To choose k ordered items from a set A, with replacement, is the same as to choose an element of A^k. We have $|A^k| = |A|^k$. Similarly $|A \times B \times C| = |A| \cdot |B| \cdot |C|$.

2. The number of maps $f : A \to B$ is $|B|^{|A|}$.

3. The number of ways to put the elements of A in order is the same as the number of bijective maps $f : A \to A$. It is given by $|A|!$.

4. The number of ways to choose k different elements of A in order, $a_1, \ldots, a_k$, is

$$(n)_k = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!},$$

   where $n = |A|$.

5. The number of injective maps $f : B \to A$ is $(n)_k$ where $k = |B|$ and $n = |A|$.

6. The subsets $\mathcal{P}(A)$ of A are the same as the maps $f : A \to 2$. Thus $|\mathcal{P}(A)| = 2^{|A|}$.

7. *Binomial coefficients.* The number of k element subsets of A is given by
$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{(n)_k}{k!}.$$
This second formula has a natural meaning: it is the number of *ordered* sets with k different elements, divided by the number of orderings of a given set. It is the most efficient method for computation. Also it is remarkable that this number is an *integer*; it can *always* be computed by cancellation. E.g.
$$\binom{10}{5} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{10}{2 \cdot 5} \cdot \frac{9}{3} \cdot \frac{8}{4} \cdot 7 \cdot 6 = 6 \cdot 42 = 252.$$

8. These binomial coefficients appear in the formula
$$(a+b)^n = \sum_0^n \binom{n}{k} a^k b^{n-k}.$$
Setting $a = b = 1$ shows that $|\mathcal{P}(A)|$ is the sum of the number of k-element subsets of A for $0 \le k \le n$.

   From the point of view of group theory, S_n acts transitively on the k-element subsets of A, with the stabilizer of a particular subset isomorphic to $S_k \times S_{n-k}$.

9. The number of ways to assign n people to teams of size $k_1, \ldots, k_s$ with $\sum k_i = n$ is
$$\binom{n}{k_1\, k_2\, \cdots k_s} = \frac{n!}{k_1! \cdots k_s!}.$$

Walks on a chessboard. The number of shortest walks that start at one corner and end at the opposite corner is $\binom{14}{7}$.

Birthdays revisited. The number of possible birthdays for k people is 365^k, while the number of possible *distinct* birthdays is $(365)_k$. So the probability that no matching birthdays occur is
$$p = \frac{(365)_k}{365^k} = \frac{k!}{365^k}\binom{365}{k}.$$

Senators on a committee. The probability p that Idaho is represented on a committee of 50 senators satisfies

$$1 - p = q = \binom{98}{50} \Big/ \binom{100}{50} = \frac{50 \cdot 49}{100 \cdot 99} = 0.247475.$$

So there is more than a 3/4 chance. Why is it more than 3/4? If there first senator is *not* on the committee, then the chance that the second one *is* rises to 50/99. So

$$p = 1/2 + 1/2(50/99) = 0.752525\ldots$$

(This is an example of conditional probabilities.) Note that this calculation is much simpler than the calculation of $\binom{100}{50}$, a 30 digit number.

The probability that *all* states are represented is

$$p = 2^{50} \Big/ \binom{100}{50} \approx 1.11 \times 10^{-14}.$$

Poker. The number of poker hands is $\binom{52}{5}$, about 2.5 million. Of these, only 4 give a royal flush. The odds of being dealt a royal flush are worse than 600,000 to one.

The probability p of a flush can be found by noting that there are $\binom{13}{5}$ hands that are all spades; thus

$$p = 4\binom{13}{5} \Big/ \binom{52}{5} = \frac{33}{16660} = 0.198\%.$$

Note that some of these flushes may have higher values, e.g. a royal flush is included in the count above.

The probability p of *just a pair* (not two pair, or three of a kind, or a full house) is also not too hard to compute. There are 13 possible values for the pair. The pair also determines 2 suits, which can have $\binom{4}{2}$ possible values. The remaining cards must represent different values, so there are $\binom{12}{3}$ choices for them. Once they are put in numerical order, they give a list of 3 suits, with 4^3 possibilities. Thus the number of hands with just a pair is

$$N = 13\binom{4}{2}\binom{12}{3}4^3 = 1,098,240,$$

and the probability is $p = N/\binom{52}{5}$ which gives about 42%. There is almost a 50% chance of having *at least* a pair.

Bridge. The number of bridge hands, $N = \binom{52}{13}$, is more than half a trillion. The probability of a perfect hand (all one suit) is $4/N$, or about 158 billion

to one. A typical report of such an occurrence appears in the Sharon Herald (Pennsylvania), 16 August 2010. Is this plausible? (Inadequate shuffling may play a role.) Note: to win with such a hand, you still need to win the bidding and have the lead.

There are 263,333 royal flushes for every perfect hand of bridge.

Flags. The number of ways to fly k flags on n poles (where the vertical order of the flags matters) is $N = n(n+1)\cdots(n+k-1)$. This is because there are n positions for the first flag but $n+1$ for the second, because it can go above or below the first.

If the flags all have the same color, there are

$$\frac{N}{k!} = \binom{n+k-1}{k} = \binom{n+k-1}{n-1}$$

arrangements. This is the same as the number of monomials of degree k in n variables. One can imagine choosing $n-1$ moments for transition in a product of k elements. Until the first divider is reached, the terms in the product represent the variable x_1; then x_2; etc.

Three types of occupation statistics. There are three types of common sample spaces or models for k particles occupying n states, or k balls dropping into n urns, etc. In all these examples, points in the sample space have equal probability; but the sample spaces are different.

Maxwell–Boltzmann statistics. The balls are numbered $1, \dots k$ and the sample space is the set of sequences $(a_1, \dots, a_k)$ with $1 \le a_i \le n$. Thus S is the set of maps $f : k \to n$. With $n = 365$, this is used to model to birthdays occurring in class of size k.

Bose–Einstein statistics. The balls are indistinguishable. Thus the sample space is just the number of balls in each urn: the set of $(k_1, \dots, k_n)$ such that $\sum k_i = k$. Equivalently S is the quotient of the set of maps $f : k \to n$ by the action of S_k.

Fermi–Dirac statistics. No more than one ball can occupy a given urn, *and* the balls are indistinguishable. Thus the sample space S is just the collection of subsets $A \subset n$ such that $|A| = k$; it satisfies $|S| = \binom{n}{k}$.

Indistinguishable particles: Bosons. How many ways can k identical particles occupy n different cells (e.g. energy levels)? This is the same as flying identical flags; the number of ways is the number of solutions to

$$k_1 + \cdots + k_n = k$$

with $k_i \ge 0$, and is given by $\binom{n+k-1}{k}$ as we have seen.

In physics, k photons arrange themselves so all such configurations are *equally likely*. For example, when $n = k = 2$ the arrangements $2 + 0$, $1 + 1$ and $0 + 2$ each have probability $1/3$. This is called *Bose–Einstein statistics*.

This is *different* from Maxwell-Boltzmann statistics, which are modeled on assigning 2 photons, one at a time, to 2 energy levels. In the latter case $1 + 1$ has probability $1/2$. Particles like photons (called bosons) cannot be distinguished from one another, so this partition counts as only *one case* physically (it does not make sense to say 'the first photon went in the second slot').

Example: Runs. If we require that every cell is occupied, i.e. if we require that $k_i \geq 1$ for all i, then the number of arrangements is smaller: it is given by

$$N = \binom{k-1}{k-n}.$$

Indeed, we can 'prefill' each of the n states with a single particle; then we must add to this a distribution of $k - n$ particles by the same rule as above.

As an example, when 5 people sit at a counter of 16 seats, the n runs of free and occupied seats determine a partition $k = 16 = k_1 + \cdots + k_n$. The pattern EOEEEOEEEOEEEOEOE, for example, corresponds to $16 = 1 + 1 + 2 + 1 + 3 + 1 + 3 + 1 + 1 + 1 + 1$. Here $n = 11$, that is, there are 11 runs (the maximum possible).

Is this evidence that diners like to have spaces between them? The number of possible seating arrangements is $\binom{16}{5}$. That is, our sample space consists of injective maps from people to seats (Fermi–Dirac statistics). To get 11 runs, each diner must have an empty seat to his right and left. Thus the number of arrangements is the same as the number of partitions of the remaining $k = 11$ seats into 6 runs with $k_1 + \cdots + k_6 = 11$. Since each $k_i \geq 1$, this can be done in $\binom{10}{5}$ ways (Bose–Einstein statistics). Thus the probability of 11 runs arising by chance is

$$p = \binom{10}{5} \Big/ \binom{16}{5} = 3/52 = 0.0577\ldots$$

Misprints and exclusive particles: Fermions. Electrons, unlike photons, obey Fermi–Dirac statistics: no two can occupy the same state (by the Pauli exclusion principle). Thus the number of ways k particles can occupy n states is $\binom{n}{k}$. For example, $1 + 1$ is the *only* distribution possible for 2 electrons in 2 states. In Fermi–Dirac statistics, these distributions are all given equal weight.

When a book of length n has k misprints, these must occur in distinct positions, so misprints are sometimes modeled on fermions.

Samples: the hypergeometric distribution. Suppose there are A defects among N items. We sample n items. What is the probability of finding a defects? It's given by:

$$p_a = \binom{n}{a}\binom{N-n}{A-a}\Big/\binom{N}{A}.$$

(Imagine the N items in order; mark A at random as defective, and then sample the first n.) Note that $\sum p_a = 1$.

We have already seen a few instances of it: for example, the probability that a sample of 2 senators includes none from a committee of 50 is the case $n = 2$, $N = 100$, $A = 50$ and $a = 0$. The probability of 3 aces in a hand of bridge is

$$p_3 = \binom{13}{3}\binom{39}{1}\Big/\binom{52}{4} \approx 4\%,$$

while the probability of no aces at all is

$$p_0 = \binom{13}{0}\binom{39}{4}\Big/\binom{52}{4} \approx 30\%.$$

Single samples. The simple identity

$$\binom{N-1}{A-1} = \frac{A}{N}\binom{N}{A}$$

shows that $p_1 = A/N$. The density of defects in the total population gives the probability that any single sample is defective.

Symmetry. We note that p_a remains the same if the roles of A and n are interchanged. Indeed, we can imagine choosing 2 subsets of N, with cardinalities A and n, and then asking for the probability p_a that their overlap is a. This is given by the equivalent formula

$$p_a = \binom{N}{a}\binom{N-a}{n-a, A-a}\Big/\binom{N}{n}\binom{N}{A},$$

which is visibly symmetric in (A, n) and can be simplified to give the formula above. (Here $\binom{a}{b,c} = a!/(b!c!(a-b-c)!)$ is the number of ways of choosing *disjoint* sets of size b and c from a universe of size a.) Here is the calculation

showing agreement with the old formula. We can drop $\binom{N}{A}$ from both sides. Then we find:

$$\binom{N}{a}\binom{N-a}{n-a, A-a}\binom{N}{n}^{-1} = \frac{N!(N-a)!n!(N-n)!}{a!(N-a)!(n-a)!(A-a)!(N-n-A+a)!N!}$$
$$= \frac{n!(N-n)!}{(n-a)!a!(A-a)!(N-n-A+a)!} = \binom{n}{a}\binom{N-n}{A-a},$$

as desired.

Limiting case. If N and A go to infinity with $A/N = p$ fixed, and n, a fixed, then we find

$$p_a \to \binom{n}{a} p^a (1-p)^{n-a}.$$

This is intuitively justified by the idea that in a large population, each of n samples independently has a chance A/N of being defective. We will later study in detail this important *binomial distribution*.

To see this rigorously we use a variant of the previous identity:

$$\binom{N-1}{A} = \binom{N-1}{N-A-1} = \left(1 - \frac{A}{N}\right)\binom{N}{A}.$$

We already know

$$\binom{N-a}{A-a} \approx \left(\frac{A}{N}\right)^a \binom{N}{A},$$

and the identity above implies

$$\binom{N-n}{A-a} \approx \left(1 - \frac{A}{N}\right)^{n-a} \binom{N-a}{A-a}.$$

This yields the desired result.

Statistics and fish. The hypergeometric distribution is important in statistics: a typical problem there is to give a confident upper bound for A based solely on a.

In Lake Champlain, 1000 fish are caught and marked red. The next day another 1000 fish are caught and it is found that 50 are red. How many fish are in the lake?

Here we know $n = 1000$, $a = 50$ and $A = 1000$, but we don't know N. To make a guess we look at the value of $N \geq 1950$ which makes p_{50} as large as possible — this is the maximum likelihood estimator. The value $N = 1950$ is very unlikely — it means all fish not captured are red. A very large value

of N is also unlikely. In fact $p_a(N)$ first increases and then decreases, as can be seen by computing

$$\frac{p_a(N)}{p_a(N-1)} = \frac{(N-n)(N-A)}{(N-n+a-A)N}.$$

(The formula looks nicer with $N-1$ on the bottom than with $N+1$ on the top.) This fraction is 1 when $a/n = A/N$. So we estimate A/N by $50/1000 = 1/20$, and get $N = 20A = 20,000$.

Waiting times. A roulette table with n slots is spun until a given lucky number — say 7 — come up. What is the probability p_k that this takes more than k spins?

There are n^k possible spins and $(n-1)^k$ which avoid the number 7. So

$$p_k = \left(1 - \frac{1}{n}\right)^k.$$

This number decreases with k. The *median* number of spins, m, arises when $p_k \approx 1/2$; half the time it takes longer than m (maybe much longer), half the time shorter. If n is large then we have the approximation

$$\log 2 = \log(1/p_m) = m/n$$

and hence $m \approx n \log 2$. So for 36 slots, a median of 25.2 spins is expected. (In a real roulette table, there are 38 slots, which include 0 and 00 for the house; the payoff would only be fair if there were 36 slots.)

Similarly, if we are waiting for a student to arrive with the same birthday as Lincoln, the median waiting time is $365 \log 2 = 255$ students.

Expected value. In general if $X : S \to \mathbb{R}$ is a real-valued function, we define its *average* or *expected value* by

$$E(X) = \sum_S p(s)X(s).$$

We can also rewrite this as

$$E(X) = \sum_t tP(X = t),$$

where t ranges over the possible values of X. This is the weighted average of X, so it satisfies $\min X(s) \leq E(X) \leq \max X(s)$. Note also that $E(X)$ need not be a possible value of X. For example, if X assumes the values 0 and 1 for a coin flip, then $E(X) = 1/2$.

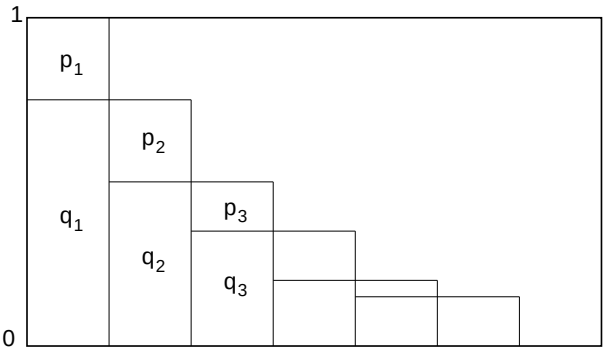

Figure 1. Formulas for waiting times: $\sum kp_k = 1 + \sum q_k$.

Expected waiting time. Suppose an event occurs at the kth trial with probability p_k, with $\sum_1^\infty p_k = 1$. Then the expected waiting time to see the event is

$$W = \sum_1^\infty kp_k = 1 + \sum_1^\infty q_k,$$

where q_k is the probability of *no success through the kth trial*.

This equality is explained in Figure 1. The heights of the rectangles are p_k and q_k respectively. The total area enclosed is $\sum kp_k$, but if we shift the graph one unit to the left, we lose a rectangle of area 1 and get a new area of $\sum q_k$.

Example: The *average* (not median) number of trials needed to see an event of probability p is $1/p$. Indeed, the probability of no success in the first k trials is q^k, where $q = 1-p$. Thus $W = 1+\sum_1^\infty q^k = 1/(1-q) = 1/p$.

A more intuitive explanation is the following. In a large number of N trials we expect to see about $k \approx pN$ events. The waiting times for each event satisfy $t_1 + \cdots + t_k = N$, so their average is $N/k \approx N/pN = 1/p$.

We will later see examples where the average waiting time is *infinite*.

Expected number of defective samples. The hypergeometric distribution gives a rather complicated formula for the probability $p_a = P(X = a)$ of finding exactly a defects. But the expected value $E(X) = \sum ap_a$ is easy to evaluate: it is $n(A/N)$. This is because $E(X) = \sum_1^n E(X_i)$ where $X_i = 1$ if the ith sample is defective and 0 otherwise. And it is easy to see that $E(X_i) = A/N$.

Median versus mean. If the grades in a class are $CCCA$, then the median grade is a C, but the average grade is between C and A — most students

are below average.

Stirling's formula. For smallish values of n, one can find $n!$ by hand, in a table or by computer, but what about for large values of n? For example, what is 1,000,000!?

Theorem II.1 (Stirling's formula) *For $n \to \infty$, we have*

$$n! \sim \sqrt{2\pi n}(n/e)^n \quad = \quad \sqrt{2\pi} n^{n+1/2} e^{-n},$$

meaning the ratio of these two quantities tends to one.

Sketch of the proof. Let us first estimate $L(n) = \log(n!) = \sum_1^n \log(k)$ by an integral. Since $(x \log x - x)' = \log x$, we have

$$L(n) \approx \int_1^{n+1} \log x \, dx = (n+1)\log(n+1) - n.$$

In fact rectangles of total area $L(n)$ fit under this integral with room to spare; we can also fit half-rectangles of total area about $(\log n)/2$. The remaining error is bounded and in fact tends to a constant; this gives

$$L(n) = (n + 1/2)\log n - n + C + o(1),$$

which gives $n! \sim e^C n^{n+1/2} e^{-n}$ for some C. The value of C will be determined in a natural way later, when we discuss the binomial distribution. ∎

Note we have used the fact that $\lim_{n\to\infty} \log(n+1) - \log(n) = 0$.

Appendix: Some calculus facts. In the above we have used some elementary facts from algebra and calculus such as:

$$\begin{aligned}
\frac{n(n+1)}{2} &= 1 + 2 + \cdots + n, \\
e^x &= 1 + x + \frac{x^2}{2!} + \cdots = \sum_0^\infty \frac{x^n}{n!}, \\
e^x &= \lim_{n\to\infty} \left(1 + \frac{x}{n}\right), \\
\log(1+x) &= x - x^2/2 + x^3/3 - \cdots = \sum_1^\infty (-1)^{n+1} x^n/n.
\end{aligned}$$

This last follows from the important formula

$$1 + x + x^2 + \cdots = \frac{1}{1-x},$$

which holds for $|x| < 1$. These formulas are frequent used in the form

$$(1-p) \approx e^{-p} \quad \text{and} \quad \log(1-p) \approx -p,$$

which are valid when p is small (the error is of size $O(p^2)$. Special cases of the formula for e^x are $(1+1/n)^n \to e$ and $(1-1/n)^n \to 1/e$. We also have

$$\log 2 = 1 - 1/2 + 1/3 - 1/4 + \cdots = 0.6931471806\ldots,$$

and

$$\int \log x \, dx = x \log x - x + C.$$

III Random Walks

One of key ideas in probability is to study not just events but *processes*, which evolve in time and are driven by forces with a random element.

The prototypical example of such a process is a *random walk* on the integers $\mathbb{Z}$.

Random walks. By a *random walk* of length n we mean a sequence of integers $(s_0, \ldots, s_n)$ such that $s_0 = 0$ and $|s_i - s_{i-1}| = 1$. The number of such walks is $N = 2^n$ and we give all walks equal probability $1/2^n$. You can imagine taking such a walk by flipping a coin at each step, to decide whether to move forward or backward.

Expected distance. We can write $s_n = \sum_1^n x_i$, where each $x_i = \pm 1$. Then the expected distance squared from the origin is:

$$E(s_n^2) = E\left(\sum x_i^2\right) + \sum_{i \neq j} E(x_i x_j) = n.$$

This yields the key insight that after n steps, one has generally wandered no more than distance about $\sqrt{n}$ from the origin.

Counting walks. The random walk ends at a point $x = s_n$. We always have $x \in [-n, n]$ and $x = n \bmod 2$. It is usual to think of the random walk as a graph through the points $(0,0)$, $(1, s_1)$, $\ldots (n, s_n) = (n, x)$.

It is useful to write

$$n = p + q \quad \text{and} \quad x = p - q.$$

(That is, $p = (n+x)/2$ and $q = (n-x)/2$.) Suppose $x = p-q$ and $n = p+q$. Then in the course of the walk we took p positive steps and q negative steps.

To describe the walk, which just need to choose which steps are positive. Thus the number of walks from $(0,0)$ to (n,x) is given by

$$N_{n,x} = \binom{p+q}{p} = \binom{p+q}{q} = \binom{n}{(n+x)/2} = \binom{n}{(n-x)/2}.$$

In particular, the number of random walks that return to the origin in $2n$ steps is given by

$$N_{2n,0} = \binom{2n}{n}.$$

Coin flips. We can think of a random walks as a record of n flips of a coin, and s_k as the *difference* between the number of heads and tails seen so far. Thus $p_x = N_{n,x}/2^n$ is the *probability* of x more heads than tails, and $N_{2n,0}$ is the probability of seeing exactly half heads, half tails in $2n$ flips.

We will later see that for fixed n, $N_{2n,x}$ approximates a bell curve, and $N_{2n,0}$ is the highest point on the curve.

Pascal's triangle. The number of paths from $(0,0)$ to (n,k) is just the sum of the number of paths to $(n-1,k-1)$ and $(n-1,k+1)$. This explains Pascal's triangle; the table of binomial coefficients is the table of the values of $N_{n,x}$, with x horizontal and n increasing down the page.

							1									
						1		1								
					1		2		1							
				1		3		3		1						
			1		4		6		4		1					
		1		5		10		10		5		1				
	1		6		15		20		15		6		1			
1		7		21		35		35		21		7		1		
1	8		28		56		70		56		28		8		1	

We can also consider walks which begin at 0, and possibly end at 0, but which must otherwise stay on the positive part of the axis. The number of such paths $P_{n,x}$ which end at x after n steps can also be computed recursively by ignoring the entries in the first column when computing the next row. Here we have only shown the columns $x \geq 0$.

1												
	1											
1		1										
	1		1									
1		2		1								
	2		3		1							
2		5		4		1						
	5		9		5		1					
5		14		14		6		1				

The ballot theorem. There is an elegant relationship between ordinary random walks, positive random walks, and random loops.

To develop this relationship, we will first show:

Theorem III.1 *The probability that a random walk from* $(0,0)$ *to* (n,x), $x>0$ *never returns to the origin is exactly its slope,* x/n.

This gives the ratio between corresponding terms in Pascal's triangle and its positive version; for example, $14/56 = 2/8 = 1/4$. (We will later see, however, that most random walks of length n have slope x/n close to zero!)

Corollary III.2 (The ballot theorem) *Suppose in an election, one candidate gets* p *votes and another* $q<p$ *votes. Then the probability that the first leads throughout the counting of ballots is* $(p-q)/(p+q)$.

The reflection principle. Let $A=(0,a)$ and $B=(n,b)$ be two points with $a,b>0$. Let $A'=(0,-a)$ be the reflection of A through the x-axis.

> *The number of paths from* A *to* B *that touch the* x*-axis is the same as the number of paths from* A' *to* B.

To see this, look at the first point P where the path from A to B hits the horizontal axis, and reflect it across the axis to obtain a path from A' to B.

Proof of Theorem III.1. The total number of walks to $(n,x)=(p+q,p-q)$ is

$$N = N_{n,x} = \binom{p+q}{p} = \binom{p+q}{q}.$$

Of these, N_+ pass through $(1,1)$ and N_- pass through $(1,-1)$. Thus

$$N = N_+ + N_- = N_{n-1,x-1} + N_{n-1,x+1}.$$

Let P be the number of walks from $(0,0)$ to (n,x) with $s_i > 0$ for $i > 0$. Then by the reflection principle,

$$P = N_+ - N_-,$$

and therefore

$$P = 2N_+ - N.$$

Since

$$N_+ = N_{n-1,x-1} = \binom{p-q-1}{p-1},$$

we have

$$\frac{N_+}{N} = \binom{p+q-1}{p-1} \Big/ \binom{p+q}{p} = \frac{p}{p+q},$$

and thus the probability of a path from $(0,0)$ to (n,x) never hitting the horizontal axis is:

$$\frac{P}{N} = \frac{2p}{p+q} - 1 = \frac{p-q}{p+q} = \frac{x}{n}.$$

■

Positive walks. We say a walk from $(0,0)$ to (n,x) is *positive* if $s_i > 0$ for $i = 1, 2, \ldots, n-1$ (since the values $s_0 = 0$ and $s_n = x$ are fixed). We also set $P_{0,0} = 1$ and $P_{n,0} = P_{n-1,1}$. The numbers $P_{n,x}$ are then entries in the one-sided Pascal's triangle.

By the ballot theorem, the number of such walks is

$$P_{x,n} = \frac{x}{n} N_{x,n}.$$

By the reflection principle, for $x > 0$, the number of such walks is:

$$P_{n,x} = N_{n-1,x-1} - N_{n-1,x+1}.$$

The first term arises because all such walks pass through $(1,1)$; the second term gets rid of those which hit the axis en route; these, by the reflection principle, are the same in number as walks from $(1,-1)$ to (n,x).

Loops. A random walk of length $2n$ which begins and ends at zero is a *loop*. The number of such random loops is given by:

$$N_{2n,0} = \binom{2n}{n} = \frac{(2n)!}{(n!)^2}.$$

By Stirling's formula, we have

$$N_{2n,n} \sim \frac{1}{\sqrt{2\pi}} \frac{(2n)^{2n+1/2} e^{-2n}}{(n^{n+1/2} e^{-n})^2} = \frac{1}{\sqrt{2\pi}} \frac{2^{2n}\sqrt{2}}{\sqrt{n}} = \frac{2^{2n}}{\sqrt{\pi n}}.$$

Thus the *probability* of a loop in $2n$ steps is

$$u_{2n} = 2^{-2n} N_{2n,n} \sim \frac{1}{\sqrt{\pi n}}.$$

We have $u_0 = 1 > u_2 > u_4 \to 0$.

Is this value for u_{2n} plausible? If we think of s_n as a random variable in $[-n, n]$ which is mostly likely of size $\sqrt{n}$, then it is reasonable that the chance that s_n is exactly zero is about $1/\sqrt{n}$.

Loops and first returns. There is an elegant relation between loops of length $2n$ and paths of length $2n$ which *never* reach the origin again. Namely we have:

Theorem III.3 *The number of paths which do not return to the origin by epoch $2n$ is the same as the number of loops of length $2n$.*

This says that the value of the middle term in a given row of Pascal's triangle is the twice the sum of the terms in the same row of the half-triangle, excluding the first.

Proof. A path which does not return to the origin is positive or negative. If positive, then at epoch $2n$ it must be at position $2k$, $k > 0$. The number of such paths, as a function of k, is given by:

$$\begin{aligned} P_{2n,2} &= N_{2n-1,1} - N_{2n-1,3}, \\ P_{2n,4} &= N_{2n-1,3} - N_{2n-1,5}, \\ P_{2n,6} &= N_{2n-1,5} - N_{2n-1,3}, \ldots \end{aligned}$$

(and for k large enough, all 3 terms above are zero). Summing these terms and multiplying by two to account for the negative paths, we obtain a total of

$$2N_{2n-1,1} = 2\binom{2n-1}{n-1}$$

paths with no return to the origin. But

$$2N_{2n-1,1} = \frac{2n}{n}\binom{2n-1}{n-1} = \binom{2n}{n} = N_{2n,0}$$

is also the number of loops. ∎

Event matching proofs. This last equality is also clear from the perspective of paths: loops of length $2n$ are the same in number as walks of one step less with $s_{2n-1} = \pm 1$.

The equality of counts suggests there should be a geometric way to turn a path with no return into a loop, and vice-versa. See Feller, Exercise III.10.7 (the argument is due to Nelson).

Corollary III.4 *The probability that the first return to the origin occurs at epoch $2n$ is:*

$$f_{2n} = u_{2n-2} - u_{2n}.$$

Proof. The set of walks with first return at epoch $2n$ is contained in the set of those with no return through epoch $2n-2$; with this set, we must exclude those with no return through epoch $2n$. ∎

Ballots and first returns. The ballot theorem also allows us to compute F_{2n}, the number of paths which make a first return to the origin at epoch $2n$; namely, we have:

$$F_{2n} = \frac{1}{2n-1}\binom{2n}{n}.$$

To see this, note that F_{2n} is number of walks which arrive at ± 1 in $2n-1$ steps without first returning to zero. Thus it is twice the number which arrive at 1. By the ballot theorem, this gives

$$F_{2n} = \frac{2}{2n-1}\binom{2n-1}{n-1} = \frac{1}{2n-1}\frac{2n}{n}\binom{2n-1}{n-1} = \frac{1}{2n-1}\binom{2n}{n}.$$

This gives

$$f_{2n} = 2^{-2n}F_{2n} = \frac{u_{2n}}{2n-1}.$$

Recurrence behavior. The probability u_{2n} of no return goes to zero like $1/\sqrt{n}$; thus we have shown:

Theorem III.5 *With probability one, every random walk returns to the origin infinitely often.*

How long do we need to wait until the random walk returns? We have $u_2 = 1/2$ so the *median* waiting time is 2 steps.

What about the *average* value of $n > 0$ such that $s_{2n} = 0$ for the first time? This average waiting time is 1+ the probability u_{2n} of failing to return by epoch $2n$; thus it is given by

$$T = 1 + \sum_{1}^{\infty} u_{2n} \sim 1 + \sum \frac{1}{\sqrt{\pi n}} = \infty.$$

Thus the returns to the origin can take a long time!

Equalization of coin flips. If we think in terms of flipping a coin, we say there is equalization at flip $2n$ if the number of heads and tails seen so far agree at that point. As one might expect, there are infinitely many equalizations, and already there is a 50% chance of seeing one on the 2nd flip. But the probability of *no equalization in the first 100 flips* is

$$u_{100} \approx \frac{1}{\sqrt{50\pi}} = 0.0797885\ldots,$$

i.e. this event occurs about once in every 12 trials.

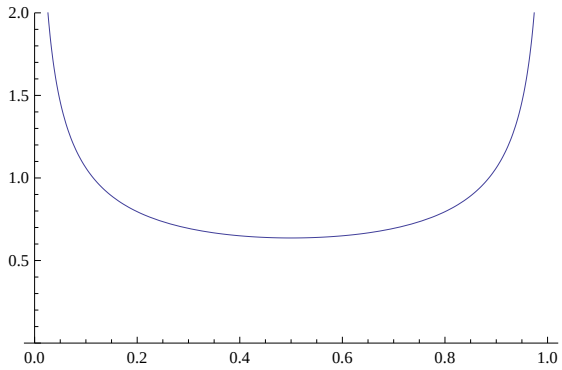

Figure 2. The arcsine density, $1/(\pi\sqrt{x(1-x)})$.

The arcsine law. Suppose in a walk of length $2n$, the latest turn to the origin occurs when $s_{2k} = 0$. Then we say $2k$ is the *latest return*.

Theorem III.6 *The probability that the latest return in a walk of length* $2n$ *occurs at epoch* $2k$ *is*

$$\alpha_{2n,2k} = u_{2k}u_{2n-2k}.$$

Remarkably, the value of α is the same for k and for $n - k$.

Proof. Such a walk consists of a loop of length $2k$ followed by a path of length $2n - 2k$ with no return. ∎

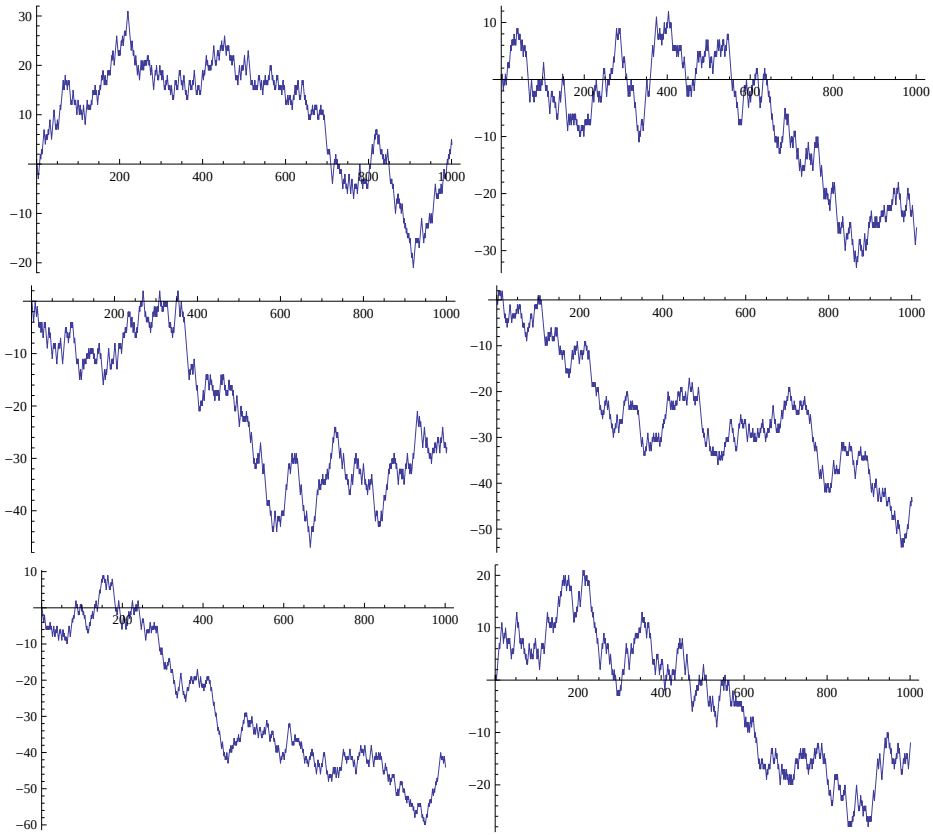

Figure 3. Typical random walks of length 1000. Time spent on the positive axis is 75%, 29%, 1%, 1%, 6%, 44%.

We already know that $u_{2k} \sim 1/\sqrt{\pi k}$, and thus

$$\alpha_{2n,2nx} \sim \frac{1}{n} \frac{1}{\pi\sqrt{x(1-x)}}.$$

Consequently, we have:

Theorem III.7 *As $n \to \infty$, the probability that the last return to the origin in a walk of length n occurs before epoch nx converges to*

$$\int_0^x \frac{dt}{\pi\sqrt{t(1-t)}} = \frac{2}{\pi} \arcsin\sqrt{x}.$$

Proof. The probability of last return before epoch nx is

$$\sum_{k=0}^{nx} \alpha_{2n,2k} = \sum_{k=0}^{nx} \alpha_{2n,2n(k/n)} \sim \sum_{k=0}^{nx} F(k/n)\,(1/n) \to \int_0^x F(t)\,dt,$$

where $F(x) = 1/(\pi\sqrt{x(1-x)})$. ∎

Remark. Geometrically, the arcsine density $dx/(\pi\sqrt{x(1-x)})$ on $[0,1]$ gives the distribution of the projection to $[0,1]$ of a random point on the circle of radius $1/2$ centered at $(1/2, 0)$.

By similar reasoning one can show:

Theorem III.8 *The percentage of time a random walk spends on the positive axis is also distributed according to the arcsine law.*

More precisely, the probability $b_{2k,2n}$ of spending exactly time $2k$ with $x > 0$ in a walk of length $2n$ is also $u_{2k}u_{2n-2k}$. Here the random walk is taken to be piecewise linear, so it spends no time at the origin. See Figure 3 for examples. Note especially that the amount of time spent on the positive axis does *not* tend to 50% as $n \to \infty$.

Sketch of the proof. The proof is by induction. One considers walks with $1 \le k \le n-1$ positive sides; then there must be a first return to the origin at some epoch $2r$, $0 < 2r < 2n$. This return gives a loop of length $2r$ followed by a path of walk of length $2n-2r$ with either $2k$ or $2k-2r$ positive sides. We thus get a recursive formula for $b_{2k,2n}$ allowing the theorem to be verified. ∎

IV Combinations of Events

We now return to the general discussion of probability theory. In this section and the next we will focus on the relations that hold between the probabilities of 2 or more events. This considerations will give us additional combinatorial tools and lead us to the notions of conditional probability and independence.

A formula for the maximum. Let a and b be real numbers. In general $\max(a, b)$ cannot be expressed as a polynomial in a and b. *However*, if a and b can only take on the values 0 and 1, then it is easy to see that

$$\max(a, b) = a + b - ab.$$

More general, we have:

Proposition IV.1 *If the numbers $a_1, \ldots, a_n$ are each either 0 or 1, then*

$$\max(a_i) = \sum a_i - \sum_{i<j} a_i a_j + \sum_{i<j<k} a_i a_j a_k + \cdots \pm a_1 a_2 \ldots a_n.$$

Proof. Suppose exactly $k \geq 1$ of the a_i's are 1. Then the terms in the sum above become

$$k - \binom{k}{2} + \binom{k}{3} + \cdots \pm \binom{k}{k}.$$

Comparing this to the binomial expansion for $0 = (1-1)^k$, we conclude that the sum above is 1. ∎

(An inductive proof is also easy to give.)

Here is another formula for the same expression:

$$\max(a_1, \ldots, a_n) = \sum_I (-1)^{|I|+1} \prod_{i \in I} a_i,$$

where I ranges over all nonempty subsets of $\{1, \ldots, n\}$.

Unions of events. Now suppose $A_1, \ldots, A_n$ are events, and $a_i(x) = 1$ if a sample x belongs to A_i, and zero otherwise. Note that $E(a_i) = P(a_i)$, $E(a_i a_j) = P(A_i A_j)$, etc. Thus we have the following useful formula:

$$\begin{aligned} P\left(\bigcup A_i\right) &= \sum_S p(x) \max(a_1(x), \ldots, a_n(x)) \\ &= \sum P(A_i) - \sum_{i<j} P(A_i A_j) + \sum_{i<j<k} P(A_i A_j A_k) + \cdots \pm P(A_1 \cdots A_n), \end{aligned}$$

where as usual AB means $A \cap B$.

Examples. The most basic case is

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

This is called the inclusion–exclusion formula; we must exclude from $P(A) + P(B)$ the points that have been counted twice, i.e. where both A and B occur. The next case is:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(AC) + P(ABC).$$

Probability of no events occurring. Here is another way to formula this result. Given events $A_1, \ldots, A_n$, let $S_1 = \sum p(A_i)$, $S_2 = \sum_{i<j} p(A_i A_j)$, etc. Let p_0 be the probability that *none* of these events occur. Then:

$$p_0 = 1 - S_1 + S_2 - \cdots \pm S_n,$$

We can rewrite this as:

$$p_0 = S_0 - S_1 + S_2 - \cdots \pm S_n,$$

where $S_0 = P(S) = 1$.

For a finite sample space, we can also think of this as a formula counting the number of points outside $\bigcup A_i$:

$$|S - \bigcup A_i| = S_0 - S_1 + S_2 + \cdots$$

where $S_0 = |S|$, $S_1 = \sum |A_i|$, $S_2 = \sum_{i<j} |A_i A_j|$, etc.

Example: typical Americans? Suppose in group of 10 Americans, 5 speak German, 4 speak French, 3 speak Italian, 2 speak German and French, 1 speaks French and Italian, and 1 speaks Italian and German. No one is trilingual. How many speak no foreign language? We have $S_0 = 10, S_1 = 5 + 4 + 3 = 12$, $S_2 = 2 + 1 + 1 = 4$, $S_3 = 0$. Thus

$$S_0 - S_1 + S_2 = 10 - 12 + 4 = 2$$

individuals speak no foreign language.

Symmetric case. Suppose the probability c_k of $P(A_{i_1} \cdots A_{i_k})$, for $i_1 < i_2 < \cdots i_k$, only depends on k. Then we have

$$S_k = \binom{n}{k} c_k,$$

and hence the probability that *none* of the events A_i occurs is given by

$$p_0 = 1 - P(A_1 \cup \cdots \cup A_n) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} c_k.$$

The mixed-up mailman revisited. What is the probability that a random permutation of n symbols has no fixed point? This is the same as the mixed-up mailman making no correct delivery on a route with n letters.

The sample space of all possible deliveries (permutations) satisfies $|S| = n!$. Let A_i be the event that the ith patron gets the right letter. Then the $n-1$ remaining letters can be delivered in $(n-1)!$ ways, so $P(A_i) = c_1 = \frac{1}{n}$. Similarly $c_2 = 1/n(n-1)$, and $c_k = 1/(n)_k$. Thus

$$S_k = \binom{n}{k} \frac{1}{(n)_k} = \frac{1}{k!},$$

and therefore

$$\begin{aligned} p_0(n) &= \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{1}{(n)_k} \\ = \sum_{0}^{n} \frac{(-1)^k}{k!} &= \frac{1}{0!} - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots \pm \frac{1}{n!}. \end{aligned}$$

Since $1/e = e^{-1} = \sum_0^\infty (-1)^k/k!$, this sum approaches $1/e$. Thus we have now rigorously verified:

> *The probability $p_0(n)$ of no correct delivery approaches $1/e$ as $n \to \infty$.*

Note that the exact probabilities here are more subtle than in the case of pesky porter, where we just had to calculate $(1 - 1/n)^n$.

Probability of exactly m correct deliveries. The number possible deliveries of n letters is $n!$. Exactly m correct delivers can occur in $\binom{n}{m}$ ways, combined with incorrect deliveries of the remaining $n - m$ letters. Thus:

$$p_m(n) = \binom{n}{m} \frac{(n-m)! p_0(n-m)}{n!} = \frac{p_0(n-m)}{m!}.$$

In particular, $p_m(n) \to 1/(m!e)$ as $n \to \infty$. Note that $\sum_0^\infty 1/(m!e) = 1$.

Probability of m events occurring. Just as

$$p_0 = 1 - S_1 + S_2 - \cdots \pm S_n,$$

we have

$$p_1 = S_1 - 2S_2 + 3S_3 - \cdots \pm nS_n,$$

and more generally

$$p_m = S_m - \binom{m+1}{m} S_{m+1} + \binom{m+2}{m} S_{m+2} + \cdots \pm \binom{n}{m} S_n.$$

Examples. The probability of exactly two of A, B, C occurring is

$$p_2 = P(AB) + P(BC) + P(AC) - 3P(ABC),$$

as can be verified with a Venn diagram.

In the delivery of n letters we have seen that $S_k = 1/k!$, and so

$$p_m = \sum_{i=0}^{n-m} (-1)^i \binom{m+i}{m} \frac{1}{(m+i)!} = \frac{1}{m!} \sum_{0}^{n-m} \frac{(-1)^i}{i!},$$

as shown before.

In the symmetric case where $S_k = \binom{n}{k} c_k$, we have

$$p_m = \binom{n}{m} \sum_{a=0}^{n-m} (-1)^a \binom{n-m}{a} c_k, \tag{IV.1}$$

since

$$\binom{m+a}{m} S_{m+a} = \binom{n}{m+a} \binom{m+a}{m} c_k = \binom{n}{m} \binom{n-m}{a} c_k.$$

The binomial relation used above is intuitively clear: both products give

$$\binom{n}{m, a}.$$

Occupancy problem; the cropduster. We now return to the study of maps $f : k \to n$, i.e. the problem of dropping k balls into n urns. What is the probability that all urns are occupied?

As a practical application, suppose we have field with n infested corn stalks. A crop dusting procedure drops k curative doses of pesticide at random. If even one infested plant remains, it will reinfect all the rest. How large should one take k to have a good chance of a complete cure? I.e. how large should the average number of multiple doses, k/n, be, if we wish to insure that each recipient gets at least one dose?

We begin with an exact solution. Let A_i be the event that the ith plant is missed. Then $P(A_i) = (n-1)^k/n^k = (1-1/n)^k$. More generally, for any subset $I \subset \{1, \ldots, n\}$, let A_I be the event that the plants in the list I receive no dose. Then $P(A_I) = (1-|I|/n)^k$. Consequently, the probability that every plant gets a dose is

$$p_0 = 1 - P(A_1 \cup \cdots \cup A_n) = \sum_{a=0}^{n} (-1)^a \binom{n}{a} \left(1 - \frac{a}{n}\right)^k.$$

Similarly, the probability that exactly m are missed is

$$p_m = \binom{n}{m} \sum_{a=0}^{n-m} (-1)^a \binom{n-m}{a} \left(1 - \frac{m+a}{n}\right)^k.$$

by equation (IV.1).

Expected number of misses. What is the *expected* number of plants that are missed? The probability of the first plant being missed after k trials is $(1-1/n)^k$. The same is true for the rest. Thus the *expected* number missed is

$$M = n(1-1/n)^k.$$

For $k = n$ and n large, this is n/e. Thus on average, more than a third of the crop is missed if we use one dose per plant. This is certainly too many misses to allow for eradication. We want the number of misses to be less than one!

More generally, for $k \gg n$ we find

$$M \approx ne^{-k/n}.$$

Thus for $k = n \log n + An$, we have

$$M = n(1-1/n)^{n \log n + An} \approx n \exp(-\log n - A) \approx \exp(-A).$$

So to reduce the expected number of misses to $1/C$, where $C \gg 1$ is our desired odds for a good outcome, we need to use

$$k = \log n + \log C$$

doses per plant. So for example with 10,000 plants, to get a million to one chance of success, it suffices to take

$$k = \log 10^5 + \log 10^6 = 11 \log 10 \approx 25,$$

i.e. to use 25,000 doses. Note that using 12,000 doses only gives about a 50% chance of success.

Bacteria. One can also think of this as a model for the use of antibiotics. Suppose we have n bacteria, and a certain dose D kills half of them. Then the expected number of survivors after a dose kD is $s = 2^{-k}n$. To get $s = 1$ we need to take $k = (\log n)/\log 2$ Then if we use a dose of $kD + aD$, the chances of a complete cure (not a single surviving bacterium) are about $1/2^a$.

V Conditional Probability

Definitions. Let (S, p) be a probability space and let $B \subset S$ be an event with $p(B) \neq 0$. We then define the *conditional probability* of A given B by

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

Put differently, from B we obtain a new probability measure on S defined by $p'(x) = p(x)/p(B)$ if $x \in B$, and $p(x) = 0$ otherwise. Then:

$$P(A|B) = P'(A) = \sum_A p'(x).$$

This new probability p' is uniquely determined by the conditions that p' is proportional to p, and that $p(x) = 0$ if $x \notin B$.

Note that $P(AB) = P(A|B)P(B)$; the chances that your instructor skis and is from Vermont is the product of the chance he is from Vermont (1/100) and the chance that a Vermonter skis (99/100). Similarly we have:

$$P(ABC) = P(A|BC)P(B|C)P(C).$$

Dice examples. When rolling a single die, let A be the event that a 2 is rolled and B be the event that an even number is rolled. Then $P(A) = 1/6$, $P(B) = 1/2$ and $P(A|B) = 1/3$. Intuitively, if you know the roll is even, then the chances of rolling a two are doubled.

Now suppose we roll two dice. Let A be the event that a total of seven is rolled, and let B be the event that we do not roll doubles. Then $P(A) = 1/6$, since for every first roll there is a unique second roll giving 7. Also $P(B) = 5/6$, since the probability of doubles is 1/6; and $P(AB) = 1/6$.

We then find $P(B|A) = 1$ — if you roll 7, you cannot roll doubles! While $P(A|B) = 1/5$ — if doubles are ruled out, you have a slightly better chance of rolling a 7.

Families with two children. In a family with 2 children, if at least one is a boy (B), what is the probability that both are boys (A)? We posit that the four possible types of families— bb, bg, gb and gg – are equally likely.

Clearly $P(B) = 3/4$, and $P(A) = 1/4$, so $P(A|B) = 1/3$. The answer is not $1/2$ — because B gives you less information than you might think; e.g. you don't know if it is the first or the second child is a boy.

(Here is a variant that gives the 'right' answer. Suppose a boy comes from a family with two children. What is the probability that his sibling is a girl? In this list of families there are 4 boys, each equally likely to be chosen. Thus $P(bb) = 1/2$ and $P(bg) = P(gb) = 1/4$. Thus the second sibling is a girl half of the time.)

Monty Hall paradox. A prize is hidden behind one of $n = 3$ doors. You guess that it is behind door one; A is the event that you are right. Clearly $P(A) = 1/3$.

Now Monty Hall opens door two, revealing that there is nothing behind it. He offers you the chance to change to door three. Should you change?

Argument 1. Let B be the event that nothing is behind door two. Clearly $P(B) = 2/3$ and $P(AB) = P(A) = 1/3$. Thus $P(A|B) = 1/2$. After Monty Hall opens door two, we know we are in case B. So it makes no difference if you switch or not.

Argument 2. The probability that you guessed right was $1/3$, so the probability you guessed wrong is $2/3$. Thus if you switch to door three, you double your chances of winning.

Argument 3. Suppose there are $n = 100$ doors. Monty Hall opens 98 of them, showing they are all empty. Then it is very likely the prize is behind the one door he refrained from opening, so of course you should switch.

Like most paradoxes, the issue here is that the problem is not well defined. Here are three ways to make it precise. We only need to determine $P(A|B)$.

(i) Monty Hall opens $n - 2$ of the remaining doors *at random*. In this case, you gain information from the outcome of the experiment, because he might have opened a door to reveal the prize. In the case of $n = 100$ doors, this is in fact very likely, unless you have already chosen the right door. So $P(A) = 1/100$ rises to $P(A|B) = 1/2$, and it makes no difference if you switch.

(ii) Monty Hall knows where the prize is, and he makes sure to open doors that do not reveal the prize, but otherwise he opens $n - 2$ doors at random. This choice reveals nothing about the status of the original door. In this case $P(A) = P(A|B) < 1/2$, and hence you should switch.

(iii) Monty Hall only opens $n-2$ doors when you guess right the first time. In this case $P(A|B) = 1$, so you should definitely not switch.

Stratified populations. Let $S = \sqcup_0^n H_i$, with $P(H_i) = p_i$. Then we have, for any event A,

$$P(A) = \sum P(A|H_i)P(H_i) = \sum p_i P(A|H_i).$$

We can also try to deduce from event A what the chances are that our sample comes from stratum i. Note that

$$\sum_i P(H_i|A) = P(S|A) = 1.$$

The individual probabilities are given by

$$P(H_i|A) = \frac{P(AH_i)}{P(A)} = \frac{P(A|H_i)P(H_i)}{\sum_j P(A|H_j)P(H_j)}.$$

For example, let S be the set of all families, and H_i those with i children. Within H_i we assume all combinations of boys and girls are equally likely. Let A be the event 'the family has no boys'. What is $P(A)$? Clearly we have $P(A|H_i) = 2^{-i}$, so

$$P(A) = \sum_0^\infty 2^{-i} P(H_i).$$

Now, suppose a family has no boys. What are the chances that it consists of just one child (a single girl)? Setting $p_i = P(H_i)$, the answer is:

$$P(H_1|A) = \frac{2^{-1}p_1}{\sum 2^{-i}p_i}.$$

To make these examples more concrete, suppose that

$$p_i = P(H_i)(1-\alpha)\alpha^i$$

for some α with $0 < \alpha < 1$. Then we find

$$P(A) = (1-\alpha)\sum(\alpha/2)^i = \frac{1-\alpha}{1-\alpha/2}.$$

When α is small, most families have no children, so $P(A)$ is close to one. When α is close to one, most families are large and so $P(A)$ is small: $P(A) \approx 2(1-\alpha)$.

To compute $P(H_1|A)$ using the formula above, only the ratios between probabilities p_i are important, so we can forget the normalizing factor of $(1-\alpha)$. Thus:

$$P(H_k|A) = \frac{2^{-k}\alpha^k}{\sum_i 2^{-i}\alpha^i} = (\alpha/2)^k(1-\alpha/2).$$

When α is close to one, we have $P(H_1|A) \approx 1/4$. Although there are many large families in this case, it is rare for them to have no boys. Indeed the most likely way for this to happen is for the family to have no children, since $P(H_0|A) = 1/2$.

Car insurance. Suppose 10% of Boston drivers are accident prone – they only have a 50% chance of an accident-free year. The rest, 90%, only have a 1% chance of an accident in a given year. This is an example of a stratified population.

An insurance company enrolls a random Boston driver. This driver has a probability $P(X) = 0.1$ of being accident prone. Let A_i be the event that the driver has an accident during the ith year.

$$P(A_1) = P(A_1|X)P(X) + P(A|\widetilde{X})P(\widetilde{X}) = 0.5 \cdot 0.1 + 0.01 \cdot 0.9 = 0.059.$$

Similarly, the probability of accidents in both of the first two years is:

$$P(A_1A_2) = 0.5^2 \cdot 0.1 + 0.01^2 \cdot 0.9 = 0.02509$$

Thus, the probability of the second accident, given the first, is:

$$P(A_2|A_1) = P(A_1A_2)/P(A_1) \approx 0.425.$$

Thus the second accident is much more likely than the first. Are accidents contagious from year to year? Is the first accident causing the second? No: the first accident is just good evidence that the driver is accident prone. With this evidence in hand, it is likely he will have more accidents, so his rates will go up.

Independence. The notion of *independence* is key in capturing one of our basic assumptions about coin flips, card deals, random walks and other events and processes.

A pair of events A and B are *independent* if

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

This is equivalent to $P(A|B) = P(A)$, or $P(B|A) = P(B)$. In other words, knowledge that event A has occurred has no effect on the probability of outcome B (and vice–versa).

Examples. Flips of a coin, steps of a random walk, deals of cards, birthdays of students, etc., are all tacitly assumed to be *independent* events. In many situations, independence is intuitively clear.

Drawing cards. A randomly chosen card from a deck of 52 has probability 1/4 of being a spade, probability 1/13 of being an ace, and probability $1/52 = (1/4)(1/13)$ of being the ace of spades. The rank and the suit are independent.

Choose 5 cards at random from a well-shuffled deck. Let B be the event that you choose the first 5 cards, and A the event that you choose a royal flush.

It is intuitively clear, and easy to check, that these events are independent; $P(A|B) = P(A)$. In other words, the probability of being dealt a royal flush is the same as the probability that the first 5 cards form a royal flush. It is for this reason that, in computing $P(A)$, we can reason, 'suppose the first 5 cards form a royal flush', etc.

Straight flush. Consider the event A that a poker hand is a straight, and the event B that it is a flush. What is $P(AB)$? We already calculated

$$P(B) = 4 \binom{13}{5} \Big/ \binom{52}{5},$$

so we only need to figure out $P(A|B)$. To compute $P(A|B)$, note that numerical values of the 5 cards in a given suit can be chosen in $\binom{13}{5}$ ways, but only 10 of these give a straight. Thus

$$P(A|B) = \left(10 \Big/ \binom{13}{5} \right) \approx 1/129.$$

The probability of a straight flush is therefore:

$$P(AB) = P(A|B)P(B) = 40 \Big/ \binom{52}{5} \approx 1/64,974.$$

On the other hand, the probability of a straight is

$$P(A) = 10 \cdot 4^5 \Big/ \binom{52}{5} \approx 1/253.$$

The conditional probability $P(A|B)$ is almost twice $P(A)$. Equivalently, $P(AB)$ is almost twice $P(A)P(B)$. Thus A and B are not independent events.

This is because once you have a flush, the numerical values in your hand must be distinct, so it is easier to get a straight.

Conversely, we also have $P(B|A) > P(B)$. To get a flush you need all the numerical values to be different, and this is automatic with a straight.

Accidental independence. Sometimes events are independent 'by accident'. In a family with 3 children, the events A=(there is at most one girl) and B=(there are both boys and girls) are independent. Indeed, we have $P(A) = 1/2$ since it just means there are more girls than boys; and $P(B) = 6/8$ since it means the family is not all girls or all boys; Finally $P(AB) = 3/8 = P(A)P(B)$, since AB means the family has one girl and two boys.

Independence of more than two events. For more than two events, independence means we also have

$$P(ABC) = P(A)P(B)P(C), \;\; P(ABCD) = P(A)P(B)P(C)P(D),$$

etc. This is stronger than just pairwise independence; it means, for example, that the outcomes of events A, B and C have no influence on the probability of D. Indeed, it implies

$$P(A|E) = P(E)$$

where E is any event that can be constructed from B, C, D using intersections and complements.

Here is an example that shows the difference. Let S be the set of permutations of (a, b, c) together with the lists $(a, a, a), (b, b, b)$ and (c, c, c); so $|S| = 9$. Let A_i be the event that a is in the ith position. Then $P(A_i) = 1/3$, and $P(A_iA_j) = 1/9$ (for $i \neq j$), but

$$P(A_1A_2A_3) = 1/9 \neq P(A_1)P(A_2)P(A_3) = 1/27.$$

This is because $P(A_k|A_iA_j) = 1$, i.e. once we have two a's, a third is forced.

Genetics: Hardy's Law. In humans and other animals, genes appear in pairs called alleles. In the simplest case two types of alleles A and a are possible, leading to three possible genotypes: AA, Aa and aa. (The alleles are not ordered, so Aa is the same as aA.)

A parent passes on one of its alleles, chosen at random, to its child; the genotype of the child combines alleles from both parents. For example, parents of types AA and aa can only produce descends of type Aa; while a parent of type Aa can contributes either an A or a, with equal probability, to its child.

Suppose in a population of N individuals, the proportions of types (AA, aa, Aa) are (p, q, r), with $p + q + r = 1$. Then an allele chosen at random from the whole population has probability $P = p + r/2$ of being A, and $Q = q + r/s$ of being a. But to choose an allele at random is the same as to choose a parent at random and then one of the parent's alleles.

Thus the probability that an individual in the next generation has genotype AA is P^2. Similarly its genotype is aa with probability Q^2 and Aa with probability $2PQ$. Thus in this new generation, the genotypes (AA, aa, Aa) appear with proportions $(P^2, Q^2, 2PQ)$, where $P + Q = 1$.

In the new generation's pool of alleles, the probability of selecting an A is now $P^2 + PQ = P$; similarly, the probability of selecting an a is Q. Thus these proportions persist for all further generations. This shows:

> *The genotypes AA, Aa and aa should occur with frequencies $(p^2, q^2, 2pq)$, for some $0 \leq p, q$ with $p + q = 1$.*

In particular, the distribution of the 3 genotypes depends on only one parameter.

We note that the distribution of genotypes in not at all stable in this model; if for some reason the number of A's increases, there is no restoring force that will return the population to its original state.

Product spaces. Finally we formalize the notion of 'independent experiments' and 'independent repeated trials'.

Suppose we have two probability spaces (S_1, p_1) and (S_2, p_2). We can then form a new sample space $S = S_1 \times S_2$ with a new probability measure:

$$p(x, y) = p_1(x)p_2(y).$$

Clearly $\sum p(x, y) = 1$. Now any event $A_1 \subset S_1$ determines an event $A_1' = A_1 \times S_2 \subset S$. This means the outcome of the first experiment conforms to A, and the second outcome is arbitrarily. We can do the same thing with events $A_2 \subset S_2$, by setting $A_2' = S_1 \times A_2$.

Then we find A_1' *and* A_2' *are independent*, since

$$P(A_1'A_2') = P(A_1 \times A_2) = P(A_1)P(A_2).$$

Thus (S, p) models the idea of doing two independent experiments. (For example: flipping a coin and rolling a die.)

Repeated trials. It is also useful to consider n *repeated independent trials* of the same experiment. This means we replace (S, p) with (S^n, p_n), where

$$p_n(x_1, \ldots, x_n) = p(x_1)p(x_2) \cdots p(x_n).$$

This is the same as $(S,p) \times (S,p) \times \cdots \times (S,p)$, n times.

A typical example, implicit in random walks and soon to be studied in more detail, is the case $S = \{H,T\}$ and $p(H) = p(T) = 1/2$. Then $p(x) = 1/2^n$ for each $x \in S^n$.

A more interesting example arises if $p(H) = r$ and $p(T) = 1 - r$. Then S^n models n flips of a biased coin. If $x = (x_1, \ldots, x_n)$ represents a heads and $n - a$ tails, then

$$p(x) = r^a(1-r)^{n-a},$$

since $p(x_i) = r$ when $x_i = H$ and $p(x_i) = (1-r)$ when $x_i = T$.

The excellent investment. We can invest in any one of a large number of different commodities, $S_1, \ldots, S_N$. Each year, commodity S_k has probability $p_k = k/N$ of going up. Different years behave independently, so the probability of S_k going up n years in a row is $(k/N)^n$.

Unfortunately we don't know which commodity (grain, oil, silicon chips, corn...) is S_i. So we invest in one C at random. It goes up for n years in a row. What is the probability that it will go up again next year?

The event U_n that investment C goes up n years in a row has probability

$$U_n = \sum P(C = S_k)(k/N)^n = \sum_1^N (k/N)^n (1/N) \approx \int_0^1 x^n \, dx = \frac{1}{n+2}$$

Thus

$$P(U_{n+1}|U_n) = P(U_{n+1})/P(U_n) \approx \frac{n+2}{n+1} \approx 1 - \frac{1}{n}.$$

This is Laplace's *law of succession*. In his time the earth was, by Biblical reckoning, about 5000 years = 1,826,213 days old. Laplace was ready to offer odds of 1,826,213 to one that the sun would rise tomorrow, based on its previous reliable behavior.

The proton. A less whimsical example is provided by the proton: experiments at the Super-Kamiokande water reactor give a lower bound on its half life of about 10^{33} years.

Coin flips. Why doesn't the same argument show that a coin which comes up 10 times in a row is very likely to come up heads 11 times as well? It would if the coin were chosen at random with a uniform bias in $[0,1]$. But a real coin, unlike the commodities above, has no memory; as a result, we find $P(U_{n+1}|U_n) = 1/2$.

In the example above, on the other hand, the outcome of the first n years is strong evidence that we have chosen a good investment. It is predicated on the assumption that a *good investment exists* — i.e. that there are *some* commodities which go up in value almost 100% of the time.

VI The Binomial and Poisson Distributions

We now turn to two of the most important general distributions in probability theory: the binomial and Poisson distributions.

The first has a simple origin from independent processes with two outcomes. The second is more subtle but in some ways even more ubiquitous.

Binomial distribution. First we explain the notion of distribution. Let $X : S \to \mathbb{Z}$ be a random variable that can only take integer values. Then its *distribution* is given by the sequence $p_k = P(X = k) \geq 0$. We have $\sum p_k = 1$; this sequence describes how the values of X are spread out or distributed among their possibilities.

For example, the *uniform distribution* on $\{1, \ldots, n\}$ is given by $p_k = 1/n$ (and $p_k = 0$ for $k < 0$ or $k > n$).

The distribution of a *single step* in a random walk is $p_1 = 1/2$, $p_{-1} = 1/2$.

Repeated experiments. Now consider a simple experiment with two outcomes: success, with probability p, and failure, with probability $q = (1 - p)$. Such an experiment is called a *Bernoulli trial*.

Let S_n be the number of successes in n independent trials. Then S_n is distributed according to the *binomial distribution*:

$$b_k = b(k; n, p) = \binom{n}{k} p^k q^{n-k}.$$

The binomial coefficient accounts for the successful trials define a subset A of $\{1, \ldots, n\}$ of cardinality k, which can be chosen in $\binom{n}{k}$ ways. By independence, the probability of each one of these events is $p^k q^{n-k}$.

Note that the *binomial formula* guarantees that $\sum_0^n b_k = 1$.

Random walks. For a random walk we have $p = q = 1/2$. The number of positive steps taken by epoch n is given by $P_n = (n + S_n)/2$. Its distribution is given by the important special case:

$$b_k = 2^{-n} \binom{n}{k}.$$

The case $p \neq q$ corresponds to a biased random walk, or an unfair coin — or simply a general event, whose probability need not be 1/2.

Why does email always crash? Suppose 6000 Harvard students check their email, day and night, on average once per hour, and checking mail takes 5 minutes. Then the probability that a given student is checking his email is $p = 1/12$. The expected number of simultaneous email checks is 500, and the probability of k checks is $b_k = \binom{n}{k} p^k q^{n-k}$.

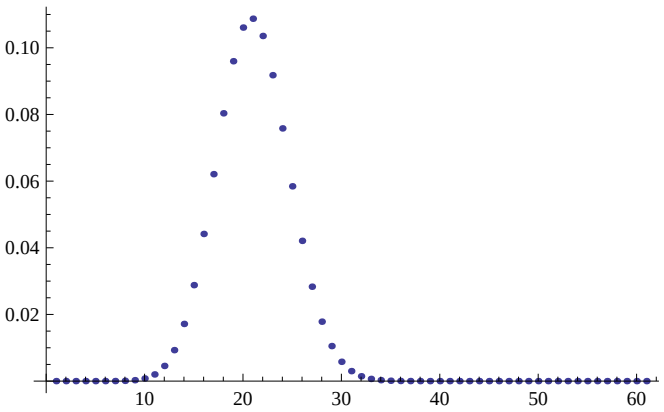

Figure 4. The binomial distribution for $n = 60$ trials, with $p = 1/3$.

The University installs 600 email servers to meet this demand. Then the probability that demand is exceeded at a given moment is

$$\sum_{k>600} b_k \approx 1/330,000.$$

If all email sessions begin at either $0, 5, 10, 15, \ldots 50$ or 55 minutes after the hour, then an overload occurs only about once every $5 \cdot 330,000$ minutes, or once every 3.14 years.

We note that with 550 servers, an overload occurs once every 7.5 hours.

Expected value. What is the expected number of successes in n Bernoulli trials with probability p each? Clearly $S_n = X_1 + \cdots X_n$ where $X_i = 1$ with probability p and 0 otherwise, so

$$E(S_n) = \sum_{1}^{n} E(X_i) = np.$$

The maximum term. It is clear that $\binom{n}{k}$ increases until $k = n/2$, then decreases, because

$$\binom{n}{k+1} = \frac{n-k}{k+1}\binom{n}{k}.$$

Thus the event of $n/2$ successes is favored because it can happen in the most ways. For fair coins this is the most likely outcome.

For an unfair coin, there is a tradeoff. If $p > q$ then events with more heads than tails are more likely, but the number of ways to achieve such

an event is smaller. Indeed, the probability of an individual event $p^k q^{n-k}$ changes by the factor p/q each time k is increased by one. Thus the ratio of successive terms in the general binomial distribution is:

$$\frac{b_{k+1}}{b_k} = \frac{n-k}{k+1}\frac{p}{q}.$$

We have only changed the fair case by a constant factor, so again the distribution is unimodal: first it goes up, then it goes down. Now the middle term occurs when

$$k \approx np.$$

(Then $(n-k)/(k+1) \approx (1-p)/p = q/p$.) This shows:

Theorem VI.1 *The most likely number of successes in n independent Bernoulli trials is $k = np$.*

Clustering near the center. It is clear from a picture (see Figure 4) that b_k is large only when k is quite close to np. (We can also see a bell curve emerging — this will be discussed soon!)

To get a feel for the size of b_k, let us consider the case $n = 2m$, $p = q = 1/2$, so $k = m$ is the middle term. Then $p^k q^{n-k} = 2^{-n}$ is the same for all k, so we need only understand how $\binom{n}{k}$ varies. Since the distribution is symmetric around the middle term $k = m$, it suffices to understand what happens for $k > m$. Here we find, for $k = m + r$,

$$\binom{n}{k} = \binom{2m}{m+r} = \binom{2m}{m} \cdot \frac{m}{m+1} \cdot \frac{m-1}{m+2} \cdots \frac{m-r+1}{m+r}.$$

This shows

$$b_{m+r}\,/b_m = \left(1 - \frac{1}{m+1}\right)\left(1 - \frac{3}{m+2}\right) \cdots \left(1 - \frac{2r-1}{m+r}\right).$$

Now we can make a very simple estimate of the probability that $S_n/n > 1/2 + s$. Set $r = sn$. We want to get an upper bound for

$$P(S_n > m + r) = \sum_{k=m+r}^{n} b_k.$$

First we note that $b_{m+r} \leq 1/r$, since $b_{m+1} \geq \cdots \geq b_{m+r}$ and the sum of these terms is at most one. Next we note that the terms in the sum satisfy

$$\frac{b_{k+1}}{b_k} \leq 1 - \frac{2r-1}{m+1}.$$

So by comparison to a geometric series, we get

$$\sum_{k=m+r}^{n} b_k \leq b_{m+r} \frac{m+r}{2r-1} \leq \frac{m+r}{r(2r-1)} \leq \frac{1}{2ns^2}.$$

Since the right-hand side goes to zero as $n \to \infty$, and since b_k is symmetric about $k = m$, we have shown:

Theorem VI.2 *For a fair coin, the number of successes in n trials satisfies, for each $s > 0$,*

$$P(|S_n/n - 1/2| > s) \to 0$$

as $n \to \infty$.

This is the simplest example of the *Law of Large Numbers*. It says that for n large, the probability is high of finding almost exactly 50% successes.

The argument same argument applies with $p \neq 1/2$, setting $m = pn$, to yield:

Theorem VI.3 *For general p, and for each $s > 0$, the average number of successes satisfies*

$$P(|S_n/n - p| > s) \to 0$$

as $n \to \infty$.

Binomial waiting times. Fix p and a desired number of successes r of a sequence of Bernoulli trials. We let f_k denote the probability that the rth success occurs at epoch $n = r + k$; in other words, that the rth success is preceded by exactly k failures. If we imagine forming a collection of successes of size r, then f_0 is the probability that our collection is complete after r steps; f_1, after $r + 1$ steps, etc.

Thus the distribution of waiting times W for completing our collection is given by $P(W = r + k) = f_k$. Explicitly, we have

$$f_k = \binom{r+k-1}{k} p^r q^k.$$

This is because we must distribute k failures among the first $r + k - 1$ trials; trial $r + k$ must always be a success.

A more elegant formula can be given if we use the convention

$$\binom{-r}{k} = \frac{(-r)(-r-1)\cdots(-r-k+1)}{k!}.$$

(This just means for integral $k > 0$, we regard $\binom{r}{k}$ as a polynomial in r; it then makes sense for any value of r. With this convention, we have:

$$f_k = \binom{-r}{k} p^r (-q)^k.$$

We also have, by the binomial theorem,

$$\sum f_k = p^r \sum_0^\infty \binom{-r}{k} (-q)^k = p^r (1-q)^{-r} = 1.$$

Thus shows we never have to wait forever.

Expected waiting time. The expected waiting time to complete r successes is $E(W) = r + \sum_0^\infty k f_k$. The latter sum can be computed by applying $q(d/dq)$ to the function $\sum \binom{-r}{k}(-q)^k = (1-q)^{-r}$; the end result is:

Theorem VI.4 *The expected waiting time to obtain exactly r successes is $n = r/p$.*

This result just says that the expected number of failures k that accompany r successes satisfies $(k : r) = (q : p)$, which is intuitively plausible. For example, the expected waiting time for 3 aces in repeated rolls of a single die is 18 rolls.

As explained earlier, there is an intuitive explanation for this. Suppose we roll a die N times. Then we expect to get $N/6$ aces. This means we get 3 aces about $N/18$ times. The total waiting time for these events is N, so the average waiting time is 18.

Poisson distribution. Suppose $n \to \infty$ and at the same time $p \to 0$, in such a way that $np \to \lambda < \infty$. Remember that np is the expected number of successes. In this case, $b_k = P(S_n = k)$ converges as well, for each k.

For example, we have

$$b_0 = \binom{n}{0} q^n = (1-p)^n \approx (1-p)^{\lambda/p} \to e^{-\lambda}.$$

The case $\lambda = 1$ is the case of the pesky porter — this limiting value is the probability of zero correct deliveries of n letters to n pigeon-holes. Similarly for any fixed k, $q^{n-k} \to e^{-\lambda}$ and $\binom{n}{k} \sim n^k/k!$. Thus

$$b_k = \binom{n}{k} p^k q^{n-k} \sim \frac{n^k p^k q^n}{k!} \to e^{-\lambda} \frac{\lambda^k}{k!}.$$

This limiting distribution is called the *Poisson distribution* of density λ:

$$p_k(\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Notice that $\sum p_k = 1$. Note also that $p_k > 0$ for all k, whereas $b_k = 0$ for $k > n$.

We note that p_k is maximized when $k \approx \lambda$, since b_k is maximized when $k \approx np$.

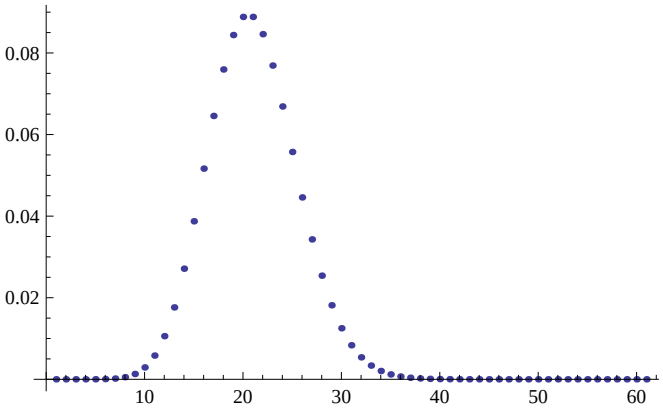

Figure 5. Poisson distribution with $\lambda = 20$.

Expected number of Poisson events. Note we have:

$$\sum kp_k(\lambda) = e^{-\lambda}\sum \frac{k\lambda^k}{k!} = e^{-\lambda}\lambda\sum \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda}\lambda e^{\lambda} = \lambda.$$

That is:

Theorem VI.5 *The expected value of a Poisson random integer $S \geq 0$ with distribution $P(S = k) = p_k(\lambda)$ is simply λ.*

This is also clear from the description of p_k as a limit of b_k, since among n Bernoulli events each with probability p, the expected number that will occur is $np \approx \lambda$.

Poisson process. Here is another way the Poisson distribution emerges. Suppose we have a way to choose a random discrete set $A \subset \mathbb{R}$. We can then count the number of points of A in a given interval I. We require that:

If I and J are disjoint, then the events $|A \cap I| = k$ and $|A \cap J| = \ell$ are independent.

The expected value of $|A \cap I| = \lambda |I|$.

Under these assumptions, we have:

Theorem VI.6 *The probability that $|A \cap I| = k$ is given by the Poisson distribution $p_k(\lambda |I|)$.*

Proof. Cut up the interval I into many subintervals of length $|I|/n$. Then the probability that one of these intervals is occupied is $p = \lambda |I|/n$. Thus the probability that k are occupied is given by the binomial coefficient $b_k(n, p)$. But the number of occupied intervals tends to $|A \cap I|$ as $n \to \infty$, and $b_k(n, p) \to p_k(\lambda |I|)$ since $np = \lambda |I|$ is constant. ∎

The bus paradox. If buses running regularly, once every 10 minutes, then the expected number of buses per hour is 6 and the expected waiting time for the next bus is 5 minutes.

But if buses run randomly and independently — so they are distributed like a Poisson process — at a rate of 6 per hour, then the expected waiting time for the next bus is 10 minutes.

This can be intuitively justified. In both cases, the waiting times $T_1, \ldots, T_n$ between n consecutive buses satisfy $(T_1 + \cdots + T_n)/n \approx 10$ minutes. But in the second case, the times T_i all have the same expectation, so $E(T_1) = 10$ minutes as well. In the first case, only T_1 is random; the remaining T_i are exactly 10 minutes.

Here is another useful calculation. Suppose a bus has just left. In the Poisson case, the probability that no bus arrives in the next minute is $p_0(1/10) = e^{-1/10} \approx 1 - 1/10$.

Thus about 1 time in 10, another bus arrives just one minute after the last one. By the same reasoning we have find:

Theorem VI.7 *The waiting time T between two Poisson events with density λ is exponentially distributed, with*

$$P(T > t) = p_0(\lambda t) = \exp(-\lambda t).$$

Since $\exp(-3) \approx 5\%$, one time out of 20 you should expect to wait more than half an hour for the bus. And by the way, after waiting 29 minutes, the *expected* remaining wait is still 10 minutes!

Clustering. We will later see that the waiting times T_i are example of exponentially distributed random variables. For the moment, we remark that since sometimes $T_i \gg 10$, it must also be common to have $T_i \ll 10$. This results in an apparent *clustering* of bus arrivals (or clicks of a Geiger counter).

Poisson distribution in other spaces. One can also discuss a Poisson process in $\mathbb{R}^2$. In this case, for any region U, the probability that $|A \cap U| = k$ is $p_k(\lambda \operatorname{area}(U))$. Rocket hits on London during World War II approximately obeyed a Poisson distribution; this motif appears in Pynchon's *Gravity's Rainbow*.

The multinomial distribution. Suppose we have an experiment with s possible outcomes, each with the same probability $p = 1/s$. Then after n trials, we obtain a partition

$$n = n_1 + n_2 + \cdots + n_s,$$

where n_i is the number of times we obtained the outcome i. The probability of this partition occurring is:

$$s^{-n} \frac{n!}{n_1! n_2! \cdots n_s!}.$$

For example, the probability that $7n$ people have birthdays equally distributed over the 7 days of the week is:

$$p_n = 7^{-7n} \frac{(7n!)}{(n!)^7}.$$

This is even rarer than getting an equal number of heads and tails in $2n$ coin flips. In fact, by Stirling's formula we have

$$p_n \sim 7^{-7n} (2\pi)^{-3} (7n)^{7n+1/2} n^{-7n-7/2} = 7^{1/2} (2\pi)^{-3} n^{-3}.$$

For coin flips we had $p_n \sim 2^{1/2} (2\pi)^{-1/2} n^{-1/2} = 1/\sqrt{\pi n}$. A related calculation will emerge when we study random walks in $\mathbb{Z}^d$.

If probabilities $p_1, \ldots, p_s$ are attributed to the different outcomes, then in the formula above s^{-n} should be replaced by $p_1^{n_1} \cdots p_s^{n_s}$.

VII Normal Approximation

We now turn to the remarkable *universal distribution* which arises from repeated trials of independent experiments. This is given by the *normal*

density function or bell curve (*courbe de cloche*):

$$\mathfrak{n}(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

Variance and standard deviation. Let X be a random variable with $E(X) = 0$. Then the size of the fluctuations of X can be measured by its *variance*,

$$\mathrm{Var}(X) = E(X^2).$$

We also have $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$. To recover homogeneity, we define the *standard deviation* by

$$\sigma(X) = \sqrt{\mathrm{Var}(X)};$$

then $\sigma(aX) = |a|\sigma(X)$.

Finally, if $m = E(X) \neq 0$, we define the variance of X to be that of the mean zero variable $X - m$. In other words,

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 \geq 0.$$

The standard deviation is defined as before.

Proposition VII.1 *We have* $\mathrm{Var}(X) = 0$ *iff* X *is constant.*

Thus the variance is one measure of the fluctuations of X.

Independence. The bell curve will emerge from sums of independent random variables. We always have $E(X+Y) = E(X)+E(Y)$ and $E(aX) = aE(X)$, but it can be hard to predict $E(XY)$. Nevertheless we have:

Theorem VII.2 *If* X *and* Y *are independent, then* $E(XY) = E(X)E(Y)$.

Proof. We are essentially integrating over a product sample space, $S \times T$. More precise, if s and t run through the possible values for X and Y respectively, then

$$\begin{aligned} E(XY) &= \sum_{s,t} stP(X = s)P(Y = t) \\ &= \Big(\sum sP(X = s)\Big)\Big(\sum tP(Y = t)\Big) = E(X)E(Y). \end{aligned}$$

∎

Corollary VII.3 *If X and Y are independent, then* $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof. We may assume $E(X) = E(Y) = 0$, in which case $E(XY) = 0$ and hence $E((X+Y)^2) = E(X) + E(Y)$. ∎

This formula is reminiscent of the Pythagorean formula; indeed, in a suitable vector space, the vectors X and Y are orthogonal.

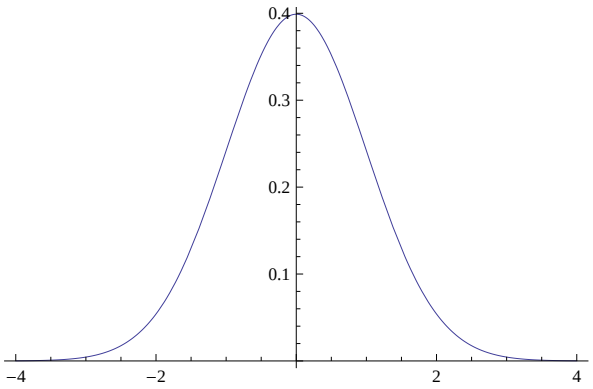

Figure 6. The normal distribution of mean zero and standard deviation one.

Standard deviation of Bernoulli sums. Consider the easy case of the random variable $S_n = \sum_1^n X_i$, where X_i are independent variables, $X_i = 1$ with probability p and 0 with probability $q = 1 - p$. Then $X_i^2 = X_i$, so $E(X_i) = E(X_i^2) = p$ and hence

$$\mathrm{Var}(X_i) = p - p^2 = pq \quad \text{and} \quad \sigma(X_i) = \sqrt{pq}.$$

Consequently

$$\mathrm{Var}(S_n) = npq \quad \text{and} \quad \sigma(S_n) = \sqrt{npq}.$$

The normal random variable. We now turn to the normal distribution and its associated random variable. To begin we establish the important fact that $\int \mathfrak{n}(x)\,dx = 1$. This explains the factor of $(2\pi)^{-1/2}$. To see this, we use the following trick:

$$\left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)^2 = \int\int e^{-(x^2+y^2)/2}\,dx\,dy = 2\pi \int_0^{\infty} e^{-r^2/2} r\,dr = 2\pi.$$

We remark that there is no simple formula for the antiderivative of $\mathfrak{n}(x)$, so we simply define the *normal distribution* by

$$N(x) = \int_{-\infty}^{x} \mathfrak{n}(y)\, dy,$$

so $N'(x) = \mathfrak{n}(x)$. We have just shown $N(x) \to 1$ as $x \to \infty$.

We can regard $\mathfrak{n}(x)$ as giving the distribution of a *random variable* X. To obtain the value of X, you choose a point under the graph of $\mathfrak{n}(x)$ at random, and the take its x-coordinate.

Put differently, we have

$$P(s \leq X \leq t) = \int_{s}^{t} \mathfrak{n}(x)\, dx = N(t) - N(s).$$

Note that $\mathfrak{n}(x) \to 0$ very rapidly as $|x| \to \infty$.

Expectation, variance and standard deviation. By symmetry, we have $E(X) = 0$. More precisely,

$$E(X) = \int x\mathfrak{n}(x)\, dx = 0$$

because $x\mathfrak{n}(x)$ is an odd function.

What about $V = E(X^2)$? This is called the *variance* of X. It can be computed by integrating x^2 against the density of the variable X: we have

$$V = \int x^2 \mathfrak{n}(x)\, dx.$$

To carry out this computation, we integrate by parts. No boundary terms appear because of the rapid decay of $\mathfrak{n}(x)$; thus we find:

$$\int x^2 e^{-x^2/2}\, dx = \int x d(-e^{-x^2}/2) = \int e^{-x^2/2}\, dx.$$

Putting back in the $(2\pi)^{-1/2}$ gives $V = 1$.

The *standard deviation* is given by $\sigma(X) = V(X)^{1/2}$, so $\sigma(X) = 1$ as well. This justifies the following terminology:

> *X is a normally distributed random variable of mean zero and standard deviation one.*

Exercise: compute $\int x^{2n}\mathfrak{n}(x)\,dx$.

Properties. The most important property of X is that its values are clustered close to its mean 0. Indeed, we have

$$\begin{aligned} P(|X|\le 1) &= N(1)-N(-1) = 68.2\%,\\ P(|X|\le 2) &= N(2)-N(-2) = 95.4\%,\\ P(|X|\le 3) &= N(3)-N(-3) = 99.7\%,\\ P(|X|\le 4) &= N(4)-N(-4) = 99.994\%, \quad \text{and}\\ P(|X|> 4) &= 2(1-N(4)) = 0.0063\%. \end{aligned}$$

The event $|X| > k\sigma$ is sometimes called a k-sigma event. For $k \ge 3$ they are very rare. One can show that

$$P(X > x) = 1 - N(x) \sim \mathfrak{n}(x)/x.$$

Indeed, integrating by parts we have

$$1-N(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-u^2/2}\,du \frac{1}{\sqrt{2\pi}}\int_x^\infty \frac{d(-e^{-u^2/2})}{u} = \frac{\mathfrak{n}(x)}{x} + \int_x^\infty \frac{\mathfrak{n}(u)\,du}{u^2},$$

and the final integral is negligible compared to the preceding term, when $x \gg 0$.

Flips of a coin and random walks. We will now show that *the final point S_n of a random walk of length n behaves like a normally distributed random variable of mean* 0 *and standard deviation* $\sqrt{n}$. What this means precisely is:

Theorem VII.4 *For any fixed $\alpha < \beta$, we have*

$$P(\alpha\sqrt{n} < S_n < \beta\sqrt{n}) \to N(\beta) - N(\alpha)$$

as $n \to \infty$.

For example, after 10,000 flips of a coin, the probability that the number of heads exceeds the number of tails by more than $100s$ is approximately by $N(-s)$ (take $\alpha = -\infty$ and $\beta = -s\sqrt{n}$). For example, a difference of 100 should occur less than 16% of the time; a difference of 200, with probability less than 2.2%.

This result is not totally unexpected, since we can write $S_n = \sum_1^n X_i$ with E_i independent and $E(X_i) = 0$, $E(X_i^2) = 1$, and thus $E(S_n) = 0$ and $E(S_n^2) = n$.

Proof. Let us rephrase this result in terms of the binomial distribution for $p = q = 1/2$. It is convenient to take $2n$ trials and for $-n \le k \le n$, set

$$a_k = b_{n+k} = 2^{-2n}\binom{2n}{n+k}.$$

This is the same as $P(S_{2n} = 2k)$. We have already seen that

$$a_0 = 2^{-2n}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}.$$

Also we have $a_{-k} = a_k$.

Now let us suppose k is fairly small compared to n, say $k \le C\sqrt{n}$. We have

$$a_k = a_0 \cdot \frac{n(n-1)\cdots(n-k+1)}{(n+1)(n+2)\cdots(n+k)} = a_0 \cdot \frac{\left(1-\frac{0}{n}\right)}{\left(1+\frac{1}{n}\right)} \cdot \frac{\left(1-\frac{1}{n}\right)}{\left(1+\frac{2}{n}\right)} \cdots \frac{\left(1-\frac{k-1}{n}\right)}{\left(1+\frac{k}{n}\right)}.$$

Approximating $1 - x$ and $1/(1 + x)$ by e^{-x}, we obtain

$$a_k \approx a_0 \exp(-(1 + 3 + 5 + \cdots 2k - 1)/n) \approx a_0 \exp(-k^2/n).$$

(Note that $(1 + 3 + \cdots 2k - 1) = k^2$, as can be seen by layering a picture of a $k \times k$ square into L-shaped strips). Here we have committed an error of $(k/n)^2$ in each of k terms, so the total multiplicative error is $1 + O(k^3/n^2) = 1 + O(1/\sqrt{n})$. Thus the ratio between the two sides of this equation goes to zero as $n \to \infty$.

If we set $x_k = k\sqrt{2/n}$, so $\Delta x_k = \sqrt{2/n}$, then $x_k^2/2 = k^2/n$ and hence we have

$$a_k \sim a_0 \exp(-x_k^2/2) \sim \frac{\exp(-x_k^2/2)}{\sqrt{\pi n}} \sim \mathfrak{n}(x_k)\,\Delta x_k.$$

Recalling that $a_k = P(S_{2n} = 2k)$, this gives:

$$\begin{aligned} P(\alpha\sqrt{2n} < S_{2n} < \beta\sqrt{2n}) &= \sum_{k=\alpha\sqrt{n/2}}^{\beta\sqrt{n/2}} a_k \sim \sum_{x_k=\alpha}^{\beta} \mathfrak{n}(x_k)\,\Delta x_k \\ &\to \int_\alpha^\beta \mathfrak{n}(x)\,dx = N(\beta) - N(\alpha) \end{aligned}$$

as claimed. ∎

Corollary VII.5 *In a random walk of length $n \gg 0$, the quantity $S_n/\sqrt{n}$ behaves like a normal random variable of mean zero and standard deviation one.*

Stirling's formula revisited. We did not actually prove that the constant in Stirling's formula is $\sqrt{2\pi}$. We can now explain how to fill this gap. We *did* show that $a_0 \sim C/\sqrt{n}$ for some constant C. On the other hand, we also know that $\sum_{-n}^{n} a_k = 1$. Using the fact that $\int \mathfrak{n}(x)\,dx = 1$, and the asymptotic expression in terms of a_0 above, we find $C = 1/\sqrt{\pi}$ and hence the constant in Stirling's formula is $\sqrt{2\pi}$.

Normal approximation for general Bernoulli trials. We now turn to the case of general Bernoulli trials, where $0 < p < 1$ and $q = 1 - p$. In this case we let S_n be the number of successes in n trials. As usual we have

$$P(S_n = k) = b_k = \binom{n}{k} p^k q^{n-k}.$$

We can think of S_n as the sum $X_1 + \cdots X_n$, where X_i are independent variables and $X_i = 1$ with probability p and $X_i = 0$ with probability q.

We now have $E(X_i) = p$ and

$$\mathrm{Var}(E_i) = E((X_i - p)^2) = q(-p)^2 + pq^2 = qp^2 + pq^2 = pq.$$

By a similar calculation we have $E(S_n) = np$ and $\mathrm{Var}(S_n) = npq$, and hence the standard deviation of S_n is given by

$$\sigma_n = \sqrt{pqn}.$$

Note that σ is *maximized* when $p = q = 1/2$; the spread of the bell curve is *less* whenever $p \neq 1/2$, because the outcomes are more predictable.

We can now state the DeMoivre–Laplace limit theorem. Its proof is similar to the one for $p = 1/2$, with just more book-keeping. We formulate two results. The first says if we express the deviation from the mean in terms of standard deviations, then we get an asymptotic formula for the binomial distribution b_k in terms of $\mathfrak{n}(x)$.

Theorem VII.6 *For $m = [np]$ and $k = o(n^{2/3})$, we have*

$$b_{m+k} \sim \mathfrak{n}(x_k)\Delta x_k,$$

where $\sigma_n = \sqrt{npq}$, $x_k = k/\sigma_n$, and $\Delta x_k = 1/\sigma_n$.

Note that we allow k to be a little bigger that $\sqrt{n}$, since in the proof we only needed $k^3/n^2 \to 0$. This small improvement will lead to the Strong Law of Large Numbers.

The second shows how the normal distribution can be used to measure the probability of deviations from the mean.

Theorem VII.7 *As $n \to \infty$ we have*

$$P(\alpha\sigma_n < S_n - np < \beta\sigma_n) \to N(\beta) - N(\alpha).$$

Put differently, if we set $X_n = (S_n - np)/\sqrt{npq}$, this gives

$$P(\alpha < X_n < \beta) \to N(\beta) - N(\alpha).$$

That means X_n behaves very much like a the standard normal random variable X of mean zero and standard deviation one. Thus when suitably normalized, all outcomes of repeated random trials are (for large n) governed by a single law, which depends on p and q only through the normalization.

How does $\sqrt{pq}$ behave? It is a nice exercise to graph $\sigma(p) = \sqrt{pq}$ as a function of $p \in [0, 1]$. One finds that $\sigma(p)$ follows a semicircle of radius $1/2$ centered at $p = 1/2$. In other words, $\sigma^2 + (p - 1/2)^2 = 1/4$.

Examples: coins and dice. Let us return to our basic examples. Consider again $n = 10{,}000$ flips of a coin. Then $p = q = 1/2$ so $\sigma = \sqrt{n}/2 = 50$. So there is a 68% chance that the number of heads S_n lies between 4,950 and 5,050; and less than a 2% chance that S_n exceeds 5,100; and S_n should exceed 5,200 only once is every 31,050 trials (i.e. even a 2% error is extremely unlikely). (Note that these answers are shifted from those for *random walks*, which have to do with the *difference* between the number of heads and tails.)

Now let us analyze $10,000$ rolls of a die. In this case $\sigma = \sqrt{n(1/6)(5/6)} = \sqrt{5n}/6 \approx 0.373\sqrt{n} \approx 37$ is somewhat less than in the case of a fair coin (but not 3 times less!). We expect an ace to come up about 1,666 times, and 68% of the time the error in this prediction is less than 37.

Median deviation. What is the *median* deviation from the mean? It turns out that $N(\alpha) - N(-\alpha) = 0.5$ for $\alpha \approx 0.6745 \approx 2/3$. (This is also the value such that $N(\alpha) = 3/4$, since $N(-\alpha) = 1 - N(\alpha)$.) Thus we have:

Theorem VII.8 *The median deviation of a normal random variable is $\approx 0.6745\sigma$. That is, the events $|X| < 0.6745\sigma$ and $|X| > 0.6745\sigma$ are equally likely, when $E(X) = 0$.*

For 10,000 flips of a fair coin, we have a median deviation of about 33 flips; for dice, about 24 rolls.

By convention a normalized IQ test has a mean of 100 and a standard deviation of 15 points. Thus about 50% of the population should have an IQ between 90 and 110. (However it is not at clear if actual IQ scores fit the normal distribution, especially for high values which occur more frequently than predicted.)

Voter polls, medical experiments, conclusions based on small samples, etc. Suppose we have a large population of N voters, and a fraction p of the population is for higher taxes. We attempt to find p by polling n members of the population; in this sample, np' are for higher taxes. How confident can we be that p' is a good approximation to p?

Assuming $N \gg n$, the polling can be modeled on n independent Bernoulli trials with probability p, and an outcome of $S_n = np'$. Unfortunately we don't know p, so how can we find the standard deviation? We observe that the largest σ can be occurs when $p = q = 1/2$, so in *any case* we have $\sigma_n \leq \sqrt{n}/2$. Consequently we have

$$\begin{aligned} P(|p' - p| \leq \alpha) &= P(|S_n - np| \leq n\alpha) \geq P(|S_n - np| \leq 2\alpha\sqrt{n}\sigma_n) \\ &\approx N(2\alpha\sqrt{n}) - N(-2\alpha\sqrt{n}). \end{aligned}$$

So for example, there is at most a 95% chance (2σ chance) that $|p - p'|$ is less than $\alpha = 1/\sqrt{n}$. So if we want an error α of at most 0.5%, we can achieve this with a sample of size $n = \alpha^{-2} = 40,000$ about 95% of the time.

Most political polls are conducted with a smaller number of voters, say $n = 1000$. In that case $\sigma \leq 1/\sqrt{1000} \approx 16$. So there is a 68% chance the poll is accurate to with $16/1000 = 1.6\%$ Errors of four times this size are very unlikely to occur (they are 4σ events). However:

> *The practical difficulty is usually to obtain a representative sample of any size.*
>
> —Feller, vol. I, p.190.

When nothing is found. If one knows or is willing to posit that p is rather small — say $p \leq 1/25$ — then a somewhat smaller sample is required, since we can replace $\sqrt{pq} \leq 1/2$ by $\sqrt{pq} \leq 1/5$. (Of course an error of a few percent may be less tolerable when the answer p is already down to a few percent itself.)

An extreme case is when a poll of n voters turns up *none* who are in favor of a tax increase. In this case simpler methods show that $p \leq C/n$ with

good confidence. Indeed, if $p > C/n$ then the probability of total failure is about $(1-p)^n \approx \exp(-C)$. We have $e^{-3} \approx 0.05$ so we have 95% confidence that $p \leq 3/n$.

This provides another approach to Laplace's estimate for the probability the sun will rise tomorrow.

Normal approximation to the Poisson distribution. We recall that the Poisson distribution $p_k(\lambda)$ describe the limiting distribution of b_k as $n \to \infty$ and $np \to \lambda$. Thus it describes a random variable S — the number of successes — with $E(S) = \lambda$. The variance of the binomial distribution is $npq \to \lambda$ as well, so we expect $\mathrm{Var}(S) = \lambda$ and $\sigma(S) = \sqrt{\lambda}$.

We can also see this directly. Let

$$f(\mu) = \exp(-\lambda) \sum_0^\infty \frac{\mu^k}{k!} = \exp(\mu - \lambda).$$

Then $f(\lambda) = \sum p_k(\lambda) = 1$. Next note that

$$\mu f'(\mu) = \exp(-\lambda) \sum_0^\infty \frac{k\mu^k}{k!} = \mu \exp(\mu - \lambda).$$

Thus $\lambda f'(\lambda) = \sum k p_k(\lambda) = E(S) = \lambda$. Similarly,

$$\left(\mu \frac{d}{d\mu}\right)^2 f = \exp(-\lambda) \sum_0^\infty \frac{k^2\mu^k}{k!} = (\mu^2 + \mu) \exp(\mu - \lambda),$$

and hence $E(S^2) = \lambda^2 + \lambda$, whence

$$\mathrm{Var}(S) = E(S^2) - E(S^2) = \lambda.$$

With these observations in place, the following is not surprising; compare Figure 5.

Theorem VII.9 *As $\lambda \to \infty$, we have*

$$P(\alpha < (S - \lambda)/\sqrt{\lambda} < \beta) \to N(\beta) - N(\alpha).$$

Similarly, we have

Theorem VII.10 *For $m = [\lambda]$ and $k = O(\lambda^{2/3})$, we have*

$$p_{m+k} \sim \mathfrak{n}(x_k)\Delta x_k,$$

where $x_k = k/\sqrt{\lambda}$ and $\Delta x_k = 1/\sqrt{\lambda}$.

Example: call centers. The remarkable feature of the Poisson distribution is that *the standard deviation is determined by the mean.* Here is an application of this principle.

The Maytag repair center handles on average 2500 calls per day. By overstaffing, it can handle an additional n calls per day. How large should n be to insure that every single call is handled 99 days out of 100?

We model the daily call volume V by the Poisson distribution (assuming a large base of mostly trouble-free customers), with expected value $\lambda = 2500$. Then the standard deviation is $\sigma = \sqrt{\lambda} = 50$. From the above, we have

$$P(V < 2500 + 50s) \approx N(s),$$

and $N(s) = 0.99$ when $s = 2.33$. Thus by adding staff to handle $n = 2.33 \cdot 50 = 117$ extra calls, we can satisfy all customers 99 days out of 100. Without the extra staffing, the support center is overloaded half the time. A 5% staff increase cuts the number of days with customer complaints by a factor of 50.

Large deviations. Using Theorem VII.6, we can now get a good bound on the probability that S_n is n^ϵ standard deviations from its mean. This bound is different from, and more elementary than, the sharper bounds found in Feller; but it will be sufficient for the Strong Law of Large Numbers. (To have an idea of the strength of this bound, note that $\exp(-n^\delta)$ tends to zero faster than $1/n^k$ for any $k > 0$, as can be checked using L'Hôpital's rule.)

Theorem VII.11 *For any $\epsilon > 0$ there is a $\delta > 0$ such that*

$$P(|S_n - np| > n^{1/2+\epsilon}) = O(\exp(-n^\delta)).$$

Proof. The probability to be estimated is bounded by nb_{np+k}, where $k = n^{1/2+\epsilon}$ (with implicit rounding to whole numbers). We may assume $0 < \epsilon < 1/6$, so Theorem VII.6 applies, i.e. we have $k = o(n^{2/3})$. Recall $\sigma_n = \sqrt{npq}$, so $x = k/\sigma_n > Cn^\epsilon > 0$ for some constant C (depending on p, q), and $\Delta x = 1/\sigma_n = O(1/\sqrt{n})$. Thus

$$nb_{np+k} \sim n\mathfrak{n}(x)\Delta x = O(\sqrt{n}\exp(-x^2/2)).$$

Since $x \gg n^\epsilon$, we have $x^2/2 \gg n^{2\epsilon}$. Thus we can choose $\delta > 0$ such that $n^\delta < x^2/4$ for all n sufficiently large. We can also drop the factor of $\sqrt{n}$, since $\exp(-n^\delta)$ tends to zero faster than any power of n. Thus $nb_{np+k} = O(\exp(-n^\delta))$ as desired. ∎

With a more careful argument, one can actually show that the probability above is asymptotic to $1 - N(x_n)$, where $x_n = n^\epsilon/\sqrt{pq}$.

VIII Unlimited Sequences of Bernoulli Trials

In this section we study infinite sequence of coin flips, dice throws and more general independent and sometimes dependent events.

Patterns. Let us start with a sequence of Bernoulli trials X_i. with $0 < p < 1$. A typical outcome is a binary sequence $(X_i) = (0, 1, 1, 0, 0, 0, 1, \ldots)$. We say a finite binary pattern $(\epsilon_1, \ldots, \epsilon_k)$ *occurs* if there is some i such that

$$(X_{i+1}, \ldots, X_{i+k}) = (\epsilon_1, \ldots, \epsilon_k).$$

Theorem VIII.1 *With probability one, every pattern occurs (infinitely often).*

In finite terms this means $p_n \to 1$, where p_n is probability that the given pattern occurs during the first n trials. One also says the pattern occurs *almost surely*.

Proof. Let A_i be the event that the given pattern occurs starting in position $i + 1$. Then A_0, A_k, A_{2k} etc. are independent events, and each has the same probability s of occurring. Since $0 < p < 1$, we have $s > 0$. So the probability that the given pattern does not occur by epoch nk is less than $P(A_0' \cdots A_n') = (1 - s)^n \to 0$. ∎

As a special case: so long as $p > 0$, there will be infinitely many successes a.s. (almost surely).

Infinite games. For example, suppose we flip a fair coin until either 3 heads come up in a row or two tails in a row. In the first case we win, and in the second case we lose. The chances of a *draw* are zero, since the patterns HHH and TT occur infinitely often (almost surely); it is just a question of which occurs first.

To calculate the probability $P(A)$ of a win, let H_1 be the event that the first flip is heads, and T_1 the event that it is tails. Then

$$P(A) = (P(A|H_1) + P(A|T_1))/2.$$

Next we claim:

$$P(A|H_1) = 1/4 + (3/4)P(A|T_1).$$

This is because either we see two more heads in a row and win (with probability 1/4), or a tail comes up, and then a winning sequence for a play beginning with a tail. Similarly we have:

$$P(A|T_1) = P(A|H_1)/2,$$

since either we immediately flip a head, or we lose. Solving these simultaneous equations, we find $P(A|H_1) = 2/5$, $P(A|T_1) = 1/5$, and hence $P(A) = 3/10$.

The Kindle Shuffle. It would not be difficult to produce a hand-held reading device that contains not just ever book every published, but every book that can possibly be written. As an add bonus, each time opened it could shuffle through all these possibilities and pick a new book which you probably haven't read before. (Engineering secret: it contains a small chip that selects letters at random until the page is full. It requires much less overhead than Borges' *Library of Babel*.)

Backgammon. The game of backgammon includes an unlimited sequence of Bernoulli trials, namely the repeated rolls of the dice. In principle, the game could go on forever, but in fact we have:

Theorem VIII.2 (M) *Backgammon terminates with probability one.*

Proof. It is easy to see that both players cannot be stuck on the bar forever. Once at least one player is completely off the, a very long (but universally bounded) run of rolls of $2-4$ will force the game to end. Such a run must eventually occur, by the second Borel–Cantelli lemma. ∎

The doubling cube also provides an interesting, simpler exercise.

Theorem VIII.3 *Player A should double if his chance of winning is $p > 1/2$. Player B should accept if $p \leq 3/4$.*

Proof. The expected number of points for player A changes from $p - q$ to $2(p - q)$ upon doubling, and this is an improvement provided $p > 1/2$. The expected point for player B change from -1 to $2(q-p)$ upon accepting, and this is an improvement if $2(p - q) < 1$ — that is, if $p < 3/4$. ∎

When $p > 3/4$ one says that player A has *lost his market*.

The Borel-Cantelli lemmas. We now consider a general infinite sequence of events $A_1, A_2, \ldots$. A *finite* sample space has only finitely many events; these results are useful only when dealing with an infinite sample space, such as an infinite random walk or an infinite sequence of Bernoulli trials. Note that the *expected number of events* which occur is $\sum P(A_i)$.

Theorem VIII.4 *If $\sum P(A_i) < \infty$, then almost surely only finitely many of these events occur.*

Proof. The probability that N or more events occur is less than $\sum_N^\infty P(A_i)$ and this tends to zero as $N \to \infty$. ∎

One might hope that $\sum P(A_i) = \infty$ implies infinitely many of the A_i occur, but this is false; for example, if $A_1 = A_2 = \cdots$ then the probability that infinitely many occur is the same as $P(A_1)$. However we *do* get a result in this direction if we have independence.

Theorem VIII.5 *If the events A_i are independent, and $\sum P(A_i) = \infty$, then almost surely infinitely many of them occur.*

Proof. It suffices to show that almost surely at least one of the events occurs, since the hypothesis holds true for $\{A_i : i > n\}$ .But the probability that *none* of the events occurs is bounded above by

$$p_n = P(A_1' \cdots A_n') = \prod_1^n (1 - P(A_i)),$$

and this tends to zero as $n \to \infty$ since $\sum P(A_i)$ diverges. ∎

Example: card matching. Consider an infinite sequence of games indexed by n. To play game n, you and your opponent each choose a number from 1 to n. When the numbers match, you win a dollar. Let A_n be the event that you win the nth game. If you choose your number at random, independently each time, then $P(A_n) = 1/n$, and by the second Borel-Cantelli lemma you will almost surely win as much money as you could ever want. (Although it will take about e^n games to win n dollars.)

Now let us modify the game so it involves 2^n numbers. Then if your opponent chooses his number at random, we have $P(A_n) = 2^{-n}$. There is no bound to the amount you can win in this modified game, but now it is almost certain that the limit of winnings as $n \to \infty$ will be finite.

Finally let us consider the *St. Petersburg game*, where chances of winning remaining 2^{-n} but the *payoff* is 2^n. Then your *expected* winnings are infinite, even though the *number of times* you will win is almost surely finite.

Example: Fermat primes and Mersenne primes. Fermat conjectured that every number of the form $F_n = 2^{2^n} + 1$ is prime. Mersenne and others considered instead the sequence $M_n = 2^n - 1$; these can only be prime when n is prime.

It is a good heuristic that the 'probability' a number of size n is prime is $1/\log n$. (In fact, the prime number theorem states the number of primes

$p \le n$ is asymptotic to $n/\log n$.) Thus the probability that M_n is prime is about $1/n$, and since $\sum 1/n$ diverges, it is conjectured that there are infinitely many Mersenne primes. (Equivalently, there are infinitely many even perfect numbers, since these all have the form $2^{p-1}(2^p - 1)$ with the second factor prime.)

On the other hand, we have $1/\log F_n \asymp 2^{-n}$, so it is conjectured that there are only finitely many Fermat primes.

There are in fact only 5 known Fermat primes: 3, 5, 17, 257, and 65537. At present there are about 47 known Mersenne primes, the largest of which is on the order of 10^{24}.

The law of large numbers. We can now formulate a stronger law of large numbers for Bernoulli trials. Before we showed that $P(|S_n/n - p| > \epsilon) \to 0$. But this could still allow fluctuations of size ϵ to occur infinitely often.

The stronger law is this:

Theorem VIII.6 *With probability one, $S_n/n \to p$.*

Proof. Fix $\alpha > 0$, and consider the events

$$A_n = (|S_n/n - p| > \alpha) = (|S_n - np| > \alpha n).$$

By Borel–Cantelli, what we need to show is that $\sum P(A_n) < \infty$ no matter what α is. Then A_n only happens finitely many times, and hence $S_n/n \to p$.

Recall that $\sigma_n = \sigma(S_n) = \sqrt{npq}$. Let us instead fix $0 < \epsilon$ and consider the event

$$B_n = (|S_n - np| > n^{1/2+\epsilon}).$$

Clearly $A_n \subset B_n$ for all $n \gg 0$, so it suffices to show $\sum P(B_n) < \infty$. But by Theorem VII.11, there is a $\delta > 0$ such that

$$P(B_n) = O(\exp(-n^\delta)).$$

Now notice that $n^\delta > 2\log n$ for all $n \gg 0$, and hence $P(B_n) = O(1/n^2)$. Since $\sum 1/n^2$ converges, so does $\sum \exp(-n^\delta)$, and the proof is complete. ■

The law of the iterated logarithm. We conclude with an elegant result due to Khintchine, which says exactly how large the deviations should get in a infinite sequence of Bernoulli trials.

Let $S_n^* = (S_n - np)/\sqrt{npq}$ be the number of standard deviations that S_n deviates from its expect value np.

Theorem VIII.7 *We have*

$$\limsup \frac{S_n^*}{\sqrt{2 \log \log n}} = 1$$

almost surely.

This means S_n^* will get bigger than $\sqrt{(2-\epsilon) \log \log n}$ infinitely often. In particular we will eventually see an event with 10 standard deviations, but we may have to wait a long time (e.g. until $2 \log \log n = 100$, or equivalently $n = \exp(\exp(50)) \approx 10^{10^{21}}$).

Here is a weaker form that is easier to prove:

Theorem VIII.8 *We have* $\limsup S_n^* = \infty$.

Proof. Choose $n_r \to \infty$ such that $\sqrt{n_{r+1}} \gg n_r$; for example, one can take $n_r = 2^{2^r}$. Let T_r^* be the deviation of the trials $(n_r + 1, \ldots, n_{r+1} - 1)$ measured in standard deviations. Now at worst, the first n_r trials were all failures; but n_r is much smaller than the standard deviation $\sqrt{n_{r+1}}$, so we still have $T_r^* = S_{n_r}^* + o(1)$. Now the events

$$A_r = (T_r^* > x)$$

are independent, and by the normal approximation they satisfy

$$P(A_r) \sim 1 - N(x) > 0,$$

so $\sum P(A_r) = +\infty$ and hence

$$\limsup S_n^* \geq \limsup T_r^* \geq x > 0.$$

Since x was arbitrary, we are done. ∎

Sketch of the proof of Theorem VIII.7. The proof is technical but straightforward. Using the Borel-Cantelli lemma, it all turns on whether or not a certain sum converges. To give the idea we will show

$$\limsup \frac{S_n^*}{\sqrt{2 \log \log n}} \leq 1.$$

The first point is that any small lead in $S_n - np$ has about a 50% probability of persisting till a later time, since about half the time S_n is above its expected value. So roughly speaking,

$$P(S_k - kp > x \text{ for some } k < n) \leq P(S_n - np > x)/2.$$

Now pick any numbers $1 < \gamma < \lambda$. We will show the limsup above is less than λ. Let $n_r = \gamma^r$, and let

$$A_r = (S_n - np \geq \lambda\sqrt{2npq \log\log n} \text{ for some } n,\ n_r \leq n \leq n_{r+1}).$$

Then we have

$$2P(A_r) \leq P(S_{n_{r+1}} - n_{r+1}p \geq \lambda\sqrt{2n_r pq \log\log n_r}).$$

The standard deviation of $S_{n_{r+1}}$ is $\sqrt{n_{r+1}pq} = \sqrt{\gamma n_r pq}$, so in terms of standard deviations the event above represents a deviation of

$$x_r = \lambda\sqrt{2\gamma^{-1}\log\log n_r} \geq \sqrt{\lambda \log\log n_r}$$

(since $\lambda > \gamma$). By the normal approximation, we then have

$$P(A_r) \leq \exp(-x_r^2/2) = \exp(-\lambda \log\log n_r) = (\log n_r)^{-\lambda} = \frac{1}{(r\log\gamma)^\lambda}.$$

Since $\sum_r 1/r^\lambda < \infty$ for any $\lambda > 1$, the first Borel-Cantelli lemma shows only finitely many of these events happen. Hence the limsup is at most one. ∎

IX Random Variables and Expectation

A *random variable* is a function $X : S \to \mathbb{R}$. In this section we give a systematic discussion of random variables.

Distribution. Let us assume S is discrete with probability function $p(s)$. Then X assumes only countably many values, say $x_1, x_2, \ldots,$. The numbers $P(X = x_i)$ give the *distribution* of X.

We have often met the case where $X : S \to \mathbb{Z}$. In this case the distribution of X is recorded by the numbers

$$p_k = P(X = k) = \sum_{s:X(s)=k} p(s).$$

We always have $\sum p_k = 1$.

Expected value. The average, mean or expected value is given by

$$E(X) = \sum_S p(s)X(s) = \sum x_i P(X = x_i).$$

In the case of an integer-valued random variable, it is given by

$$E(X) = \sum kp_k.$$

(These sums might not converge absolutely, in which case we say the expected value does not exist.) Clearly E is linear:

$$E\left(\sum a_i X_i\right) = \sum a_i E(X_i).$$

Examples.

1. For a single Bernoulli trial we have $p_0 = q$ and $p_1 = p$, and $E(X) = p$.

2. For S_n, the number of successes in n Bernoulli trials, we have

$$p_k = \binom{n}{k} p^k q^{n-k}.$$

   We have $E(S_n) = np$.

3. If X obeys a Poisson distribution then

$$p_k = e^{-\lambda} \frac{\lambda^k}{k!},$$

   and $E(X) = \lambda$.

4. The moment $X = k$ at which a random walk first returns to the origin is distributed according to:

$$p_{2n} = \frac{1}{2^{2n}(2n-1)} \binom{2n}{n}.$$

   We have $p_k = 0$ if k is odd. We have also seen then $2np_{2n} \sim \sqrt{1}\sqrt{\pi n}$, so $E(X) = +\infty$.

5. The time $X = k$ of the first success in a sequence of Bernoulli trials obeys the *geometric distribution*:

$$p_k = pq^{k-1}.$$

   This is because all the preceding events must be failures. We have $E(X) = 1/p$.

Generating functions. (See Feller, Ch. XI for more details.) When X only takes integer values $k \geq 0$, and these with probability p_k, the distribution of X is usefully recorded by the *generating function*

$$f(t) = \sum_0^\infty p_k t^k.$$

Note that $f(1) = 1$, and $f'(1) = E(X)$.

Examples.

1. For the binomial distribution for n trials with probability p, we have

$$f(t) = \sum \binom{n}{k} p^k q^{n-k} t^k = (q + tp)^n.$$

2. For the Poisson distribution with expected value λ, we have

$$f(t) = e^{-\lambda} \sum \frac{\lambda^k t^k}{k!} = \exp(\lambda t)/\exp(\lambda).$$

3. For the geometric distribution $p_k = q^{k-1}p$, we have

$$f(t) = p \sum_1^\infty q^{k-1} t^k = \frac{pt}{1 - tq}.$$

Theorem IX.1 *The generating function for a sum of independent random variables $X_1 + X_2$ is given by $f(t) = f_1(t) f_2(t)$.*

Proof. The distribution of the sum is given by

$$p_k = \sum_{i+j=k} P(X_1 = i \text{ and } X_2 = j) = \sum_{i+j=k} p_i^1 p_j^2.$$

■

Corollary IX.2 *If $S_n = \sum_1^n X_i$ is a sum of independent random variables with the same distribution, then its generating function satisfies $f_n(t) = f_1(t)^n$.*

The binomial distribution is a prime example.

Corollary IX.3 *The sum of two or more independent Poisson random variables also obeys the Poisson distribution.*

The concept of a random variable. A basic principle is that all properties of X alone are recorded by its distribution function, say p_k. That is, the sample space is irrelevant. Alternatively we could suppose $S = \mathbb{Z}$, $X(k) = k$ and $p(k) = p_k$. One can even imagine a roulette wheel of circumference 1 meter, with intervals of length p_k marked off and labeled k. Then the wheel is spun and the outcome is read off to get X.

For example, we have

$$E(\sin(X)) = \sum \sin(p_k)p_k,$$

and similarly for any other function of X. In particular $E(X)$ and $\sigma(X)$ are determined by the probability distribution p_k.

In the early days of probability theory the sample space was not explicitly mentioned, leading to logical difficulties.

Completing a collection. Each package of Pokémon cards contains 1 of N possible legendary Pokémon. How many packs do we need to buy to get all N? (Gotta catch 'em all ®.) We assume all N are equally likely with each purchase.

Let S_r be the number of packs needed to first acquire r legendaries. So $S_0 = 0$ and $S_1 = 1$, but S_r is a random variable for $r > 1$. Write $X_r = S_{r+1} - S_r$; then X_r is the waiting time for the first success in a sequence of Bernoulli trials with probability $p = (N - r)/N$. We have seen already that $E(X_r) = 1/p = N/(N - r)$. Thus

$$E(S_r) = E(X_0 + X_1 + \cdots + X_{r-1}) = N\left(\frac{1}{N} + \frac{1}{N-1} + \cdots + \frac{1}{N-r+1}\right).$$

(*Correction:* before (3.3) on p. 225 of Feller, X_r should be X_{r-1}.) To get a complete collection the expected number of trials is

$$E(S_N) = N(1 + 1/2 + 1/3 + \cdots 1/N) \sim N \log N.$$

For $N = 10$ we get $E(S_N) = 29.29$, even though $E(S_5) = 10(1/6 + 1/7 + \cdots 1/10) \approx 6.46$.

Note especially that we calculated $E(S_r)$ without every calculating the probability distribution $P(S_r = k)$. Expectations are often easier to compute than distributions.

Joint distribution. We can similarly consider two (or more random variables) X and Y. Their behavior is then governed by their *joint distribution*

$$p_{st} = P(X = s, Y = t).$$

For example, if we drop 3 balls into 3 cells, and let X be the number that land in cell one and N the number of occupied cells. Then we have 27 possible configurations, each equally likely. The row and column sums give the distributions of N and X individually.

$N\backslash X$	0	1	2	3	
1	2/27	0	0	1/27	1/9
2	6/27	6/27	6/27	0	6/9
3	0	6/27	0	0	2/9
	8/27	12/27	6/27	1/27	1

Independence. We say X and Y are independent if $P(X = s, Y = t) = P(X = s)P(Y = t)$ for all s and t. A similarly definition applies to more than two variables.

Clearly X and N above are not independent. As a second example, in a Bernoulli process let X_1 be the number of failures before the first success, and let X_2 be the number of failures between the first two successes. Then

$$P(X_1 = j, X_2 = k) = q^{j+k}p^2 = p(X_1 = j)P(X_2 = k).$$

These two events are independent.

Random numbers of repeated trials. Let N be a Poisson random variable with $E(N) = \lambda$, and let S and F denote the number of successes and failures in N Bernoulli trials with probability p. Then

$$\begin{aligned} P(S = a, F = b) &= P(N = a + b)\binom{a+b}{a}p^a q^b = e^{-\lambda}\frac{\lambda^{a+b}}{(a+b)!}\frac{(a+b)!}{a!b!}p^a q^b \\ &= \left(e^{-p\lambda}\frac{(p\lambda)^a}{a!}\right)\left(e^{-q\lambda}\frac{(q\lambda)^b}{b!}\right). \end{aligned}$$

In other words, S and F are *independent Poisson variables* with expectations λp and λq respectively.

To explain this, imagine making a huge batch of muffins with an average of λ raisins per muffin. Then the number of raisins in a given muffin is

distributed by N. Suppose on average $p\lambda$ raisins are black and $q\lambda$ are gold. Then S and F give the number of black and gold raisins. Clearly their distributions are independent and Poisson.

Covariance. We have previously seen:

Theorem IX.4 *If X and Y are independent then $E(XY) = E(X)E(Y)$ and $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.*

We can use the failure of these results to hold as *one* measure of the dependence between X and Y. Namely we define the *covariance* by

$$\operatorname{Cov}(X,Y) = E(XY) - E(X)E(Y).$$

Note that like the variance, $\operatorname{Cov}(X+a, Y+b) = \operatorname{Cov}(X,Y)$; we can always reduce to the case of variables with mean zero. It also satisfies $\operatorname{Cov}(aX, bY) = ab\operatorname{Cov}(X,Y)$ and $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$. Bilinearity is also very useful:

$$\operatorname{Cov}(X_1 + X_2, Y) = \operatorname{Cov}(X_1, Y) + \operatorname{Cov}(X_2, Y).$$

Finally note that

$$\operatorname{Var}(X) = \operatorname{Cov}(X,X).$$

Theorem IX.5 *We have $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y)$. More generally,*

$$\operatorname{Var}\left(\sum X_i\right) = \sum \operatorname{Var}(X_i) + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j).$$

Proof. We can assume mean zero; then in the first case we have

$$\operatorname{Var}(X+Y) = E((X+Y)^2) = E(X^2) + E(Y^2) + 2E(XY)$$

and the result is immediate. The general case is similar. ∎

Here is a variant of the familiar Cauchy–Schwarz inequality, $|\langle x, y\rangle|^2 \leq \|x\|^2 \cdot \|y\|^2$.

Theorem IX.6 *We have $|\operatorname{Cov}(X,Y)|^2 \leq \operatorname{Var}(X)\operatorname{Var}(Y)$.*

Proof. We may assume both sides have mean zero, and rescale so $\operatorname{Var}(X) = \operatorname{Var}(Y) = 1$. Then $0 \leq \operatorname{Var}(X-Y) = 2(\operatorname{Var}(X)^2 - \operatorname{Cov}(X,Y))$ so $\operatorname{Cov}(X,Y) \leq \operatorname{Var}(X)^2 = \operatorname{Var}(X)\operatorname{Var}(Y)$. The reverse inequality comes from the fact that $\operatorname{Var}(X+Y) \geq 0$. ∎

Correlation. A somewhat more 'natural' quantity is the *correlation coefficient*,

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

This is a dimensionless version of the covariance, i.e. $\rho(aX+b, cY+d) = \rho(X,Y)$. We also have, by the above, $|\rho(X,Y)| \leq 1$. Also $\rho(X,X) = 1$, $\rho(X,-X) = -1$ and $\rho(X,Y) = 0$ if X and Y are independent. Intuitively, if $\rho(X,Y) > 0$, then when X goes up Y also tends to go up (and vice versa).

Example. Consider our functions X and N for 3 balls dropped into 3 cells. We have $E(X) = 1$ (each ball has a 1/3 chance of landing in the first cell), and $E(N) = 19/9$ from the table above. We also see that $E(XN) = 57/27 = 19/9$. Thus:

$$\rho(X,N) = \mathrm{Cov}(X,N) = 0.$$

This shows variables can have zero correlation without being independent.

$X_2 \backslash X_1$	0	1	2	3	
0	1/27	3/27	3/27	1/27	8/27
1	3/27	6/27	3/27	0	12/27
2	3/27	3/27	0	0	6/27
3	1/27	0	0	0	1/27
	8/27	12/27	6/27	1/27	1

We can similarly examine the counts X_1 and X_2 of balls landing in cells 1 and 2 respectively. For these we find $E(X_i) = 1$, $E(X_i^2) = 5/3$, so $\mathrm{Var}(X_i) = 2/3$; and $E(X_1X_2) = 2/3$. (For the latter, note that $X_1 = X_2 = 1$ can occur in 6 ways, but $X_1 = 1, X_2 = 2$ can only occur in 3 ways.) The correlation coefficient is thus

$$\rho(X_1,X_2) = \frac{-1/3}{2/3} = -1/2.$$

More balls in cell one means less for cell two.

Covariances and dice. The covariance of the number of 1s and the number of 6s in n rolls of a die (problem IX.9(21)) is mostly easily calculated using the *bilinearity* of the covariance, the representation of each variable as a sum over trials, and the vanishing of the covariance for independent variables.

Let $A_i = 1$ when the ith roll is an ace, and zero otherwise. Define S_i similarly for a roll of six. Then

$$\mathrm{Cov}(A_i, S_i) = E(A_i S_i) - E(A_i)E(S_i) = -1/36,$$

and $\mathrm{Cov}(A_i, S_j) = 0$ for $i \neq j$ by independence. Thus

$$\mathrm{Cov}\left(\sum A_i, \sum S_i\right) = n\,\mathrm{Cov}(S_1, A_1) = -n/36.$$

We also have $\mathrm{Var}(A_i) = \mathrm{Var}(S_i) = npq = 5n/36$, so we get a correlation coefficient

$$\rho\left(\sum S_i, \sum A_i\right) = -1/5,$$

independent of n. By the same reasoning we get:

Proposition IX.7 *If (X_i, Y_i) have the same joint distribution as (X_1, Y_1) and these variables are independent for different values of i, then*

$$\rho\left(\sum_1^n X_i, \sum_1^n Y_i\right) = \rho(X_1, Y_1).$$

We remark that if A and B are Bernoulli trials with the same probability p, thought of also as $0/1$ functions, then

$$\rho(A, B) = \frac{P(AB) - p^2}{pq} = \frac{P(A|B) - P(A)}{q}.$$

For example, if A and B are mutually exclusive, then $\rho(A, B) = -p/q$. In the dice case this gives $-1/5$ as above. Note that if $P(AB) = 0$ then $P(A) + P(B) = 2p \leq 1$ and so $p \leq q$. We also note that for mean zero random variables, we have

$$\rho(X, Y) = \frac{E(XY)}{\sqrt{E(X^2)E(Y^2)}}.$$

Chebyshev's inequality. There is a very simple and yet useful way to bound the fluctuations in a random variable X of finite variance.

Theorem IX.8 *For any random variable X with $m = E(X)$ finite, we have*

$$P(|X - m| > s) \leq \mathrm{Var}(X)/s^2.$$

Proof. We may assume $E(X) = 0$, and then it is clear that

$$s^2 P(|X| > 2) \leq E(X^2) = \mathrm{Var}(X).$$

∎

This bound is sometimes quite weak, since in the case of a normal random variable with variance 1 we have $P(|X| > s) = O(\exp(-s^2/2)/s)$. On the other hand if X has a discrete distribution with $p_k \asymp 1/k^{3+\epsilon}$, then X has finite variance and

$$P(|X - m| > k) \asymp \frac{1}{k^{2+\epsilon}}.$$

So in this case Chebyshev's bound is nearly sharp.

X Law of Large Numbers

In this section we give two applications of Chebyshev's inequality. First, we generalize the strong law of large numbers beyond Bernoulli trials. The new proof is in fact easier that the one given previously, since it does not rely on the normal approximation. Second, we give an application to the combinatorics of permutations.

Theorem X.1 *The strong law of large numbers holds for any set of independent, uniformly bounded random variables X_k. That is, if $S_n = X_1 + \cdots + X_n$ and $E(S_n/n) = \mu_n$, then*

$$|S_n/n - \mu_n| \to 0$$

almost surely as $n \to \infty$.

Proof. We may assume $E(X_i) = 0$. By boundedness and independence, we have $\mathrm{Var}(S_n) = O(n)$. By Chebyshev's inequality, $P(|S_n| > \epsilon n)$ is $O(1/n)$. So if we choose a subsequence n_i along which $\sum 1/n_i < \infty$, by easy Borel-Cantelli we have $|S_{n_i}| < \epsilon n_i$ for all i sufficiently large. By boundedness, $|S_j - S_{n_i}| = O(n_{i+1} - n_i) = O(d_i)$ if $n_i < j \le n_{i+1}$. So if we can also arrange that $d_i = o(n_i)$ we are done, for then

$$\frac{|S_j|}{j} = O\left(\frac{d_i + |S_{n_i}|}{n_i}\right)$$

and both d_i/n_i and S_{n_i}/n_i tend to zero. Note that for $n_i = \lambda^i$ we do not have $d_i/n_i \to 0$. But many other choices of n_i work, e.g. if $n_i = i^2$ then $\sum 1/n_i < \infty$ and $d_i = O(i) = o(n_i)$ and we are done. ∎

This proof illustrates a key technique: we use the fact that S_n is *slowly varying* to transport a conclusion along a *sparse subsequence* of n to the whole sequence.

Here is a variant which relaxes the assumption of uniform boundedness.

Theorem X.2 *If X_i are independent, identically distributed random variables and $\mu = E(X_i)$ is finite, then $S_n/n \to \mu$ almost surely.*

In the result above, the variance of X_i is allowed to be infinite. If it is finite, however, we have:

Theorem X.3 (Central limit theorem) *If, in addition, $\sigma = \sigma(X_i)$ is finite, then*

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} < x\right) \to N(x).$$

Example: permutations. Here is a combinatorial application whose conclusions make no reference to probability.

A random permutation $\pi : n \to n$ can be constructed by first choosing $\pi(1)$, then choosing $\pi(\pi(1))$, etc. When $\pi^k(1) = 1$ we have completed a cycle; then we start over and choose $\pi(x)$, where x is the smallest integer for which π is still undefined.

Let $X_k = 1$ if a cycle is completed at the kth step (starting with $k = 0$), and 0 otherwise. Clearly $P(X_0) = 1/n$ and in general $\pi(X_k) = 1/(n-k)$. Thus the *average number of cycles* of a random permutation is given by

$$E\left(\sum X_i\right) = E(S_n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \sim \log n.$$

We also have

$$\mathrm{Var}(X_k^2) = E(X_k) - E(X_k)^2 \approx \frac{1}{n-k}$$

unless k is close to n. It is a beautiful fact that the variables X_k are independent; using this, we find that $\mathrm{Var}(S_n) \sim \log n$ as well. Then by Chebyshev, most permutations have within $O(\sqrt{\log n})$ of $\log n$ cycles. (In fact the central limit theorem holds as well.)

A related fact: a random sparse graph on n vertices has diameter about $\log n$.

Application: the median. Let $X_1, X_2, \ldots$ be an arbitrary sequence of independent, identically distributed random variables, with density $f(x)$. Their *median* is the unique value M such that

$$\int_{-\infty}^{M} f(x)\,dx = 1/2.$$

In other words, $P(X < M) = P(X > M) = 1/2$.

We claim that the median $M_n = X_{n)}$ of a finite set of trials $X_1, \ldots, X_{2n}$ converges to M almost surely, at least if $f(x) > 0$ for all x. To see this, fix $M' > M$ and let $p = P(X > M') < 1/2$. By the law of large numbers, the number of i such that $X_i > M'$ is about $pn < 1/2$ for all large enough n. Thus $M_n < M'$ for all large enough n, so $\limsup M_n \leq M$. The same reasoning shows $\liminf M_n \geq M$. We only need the theory of Bernoulli trials to make this conclusion.

XI Integral–Valued Variables. Generating Functions

See notes to Chapter IX for a brief account of some important generating functions.

XIV Random Walk and Ruin Problems

In this section we revisit the theory of random walks to address two more aspects of the theory: (i) the connection with potential theory and differential equations, and (ii) walks in higher dimensions.

Ruin and the maximum principle. Suppose we have \$1000 to gamble with at a roulette table. We bet \$1 on red every time, and quit once have made \$50. What are the chances of success?

Let us first assuming our winning chances are $p = 1/2$. Then we are simply taking a random walk on the interval $[0, 1050]$ and stopping when we first hit one of the endpoints. Starting at position $0 < n < 1050$, let s_n be the probability of success, i.e. hitting 1050 before hitting zero. Then clearly

$$s_n = (s_{n-1} + s_{n+1})/2, \tag{XIV.1}$$

where $s_0 = 0$ and $s_{1050} = 1$. One solution to this system of equations is $s_n = n/1050$. Suppose we have another solution, s'_n. Then $t_n = s_n - s'_n$ vanishes at $n = 0$ and $n = 1050$; consider any point $0 < n < 1050$ where it achieves its maximum. Then the linear equation above shows $t_n = t_{n-1} = t_{n+1}$, and hence t_n is constant, and hence zero. Thus $s_n = n/1050$ is the *only* solution satisfying the given boundary conditions. We conclude:

The chances of winning \$50 are 20/21 = 95.2% .

This is an example of a *fair game*; on average, our fortune is neither increased nor decreased, because:

Our expected gain is $50(20/21) - 1000/21 = 0$.

In fact, we could have used this fairness to give a direct calculation of the probability of success.

Expected length of the game. How long will our gambler take to finally win or be ruined? Let T_n be the expected length of the game (number of plays) starting with n dollars. Then $T_0 = 0$, $T_{1050} = 0$, and

$$T_n = 1 + (T_{n-1} + T_{n+1})/2. \tag{XIV.2}$$

The solution to this difference equation is

$$T_n = n(1050 - n).$$

The main point here is that the average of $(n-1)^2 = n^2 - 2n + 1$ and $(n+1)^2 = n^2 + 2n + 1$ is $n^2 + 1$. In particular:

> *An average of 50,000 spins are required to complete the game, starting with \$1000.*

Equally striking: if we start with \$1 and quit only if we make \$1000 or are ruined, then the avenge number of spins to complete the game is 1000. If the house has infinite resources, the expected waiting time for ruin is infinite.

The House's take. However, this is not how roulette works. In European roulette there is a green slot marked 0 that pays out to the house. Thus the probability of winning a bet on red is $p = 18/37 < 1/2$, and the probability of losing is $q = 19/37$. We now have:

$$s_n = qs_{n-1} + ps_{n+1}. \tag{XIV.3}$$

Observe that one solution is given by

$$s_n = (q/p)^n,$$

since

$$q(q/p)^{n-1} + p(q/p)^{n+1} = (q/p)^n(p+q) = (q/p)^n.$$

Another solution is given by $s_n = 1$. To satisfy the boundary conditions $s_0 = 0$, $s_{1050} = 1$, we take the linear combination

$$s_n = \frac{(q/p)^n - 1}{(q/p)^{1050} - 1}$$

Thus the chances of winning \$50 are now:

$$s_{1000}/s_{1050} \approx (p/q)^{50} = (18/19)^{50} = 6.7\%.$$

Now our expected losses are about \$933.

This games is clearly very unfair, even though they seem like small modifications of fair games. It is much better in these circumstances to make one large bet than many small ones. Betting all \$1000 at once gives an expected loss of $\$1000(q-p) \approx \$27$.

Note: in American roulette there are two green squares, and our chances of winning become $(18/20)^{50} = 0.5\%$. Our expected losses are about \$994.

Differential equations. There is a close connection between differential equations and random walks which is hinted at above. The difference equations (XIV.1), (XIV.2) and (XIV.3) are related to the differential equations $s'' = 0$, $T'' = 1$ and $s'' + \delta s' = 0$.

Higher-dimensional random walks. Let W_n be a random walk starting at the origin in some space, for example in $\mathbb{Z}^d$. Let $u_n = P(W_n = 0)$.

We say W_n is *transient* if $W_n \to \infty$ almost surely. We say it is *recurrent* if W_n returns to the origin infinitely often, almost surely.

Theorem XIV.1 *A random walk is transient if $\sum u_n < \infty$, and otherwise recurrent.*

Proof. If $\sum u_n < \infty$ then W_n only returns to the origin a finite number of times, by easy Borel–Cantelli. But then W_n must tend to infinity, otherwise it returns infinitely often to *some point*, and hence to *every point*, including the origin.

For the converse, let f_n denote the probability of a *first* return to the origin at epoch n. Then $u_0 = 1$ while for $n > 0$ we have

$$u_n = f_n u_0 + f_{n-1} u_1 + \cdots f_1 u_{n-1}.$$

So if we write $U(t) = \sum u_n t^n$ and $F(t) = \sum f_n t^n$, then $U(t) = 1 + U(t)F(t)$. Equivalently, $F(t) = 1 - 1/U(t)$.

Now suppose $U(1) = \sum u_n = \infty$. Then $F(1) = \sum f_n = 1$; but this means exactly the W_n returns to the origin with probability one. ∎

Theorem XIV.2 *A random walk W_n in $\mathbb{Z}^d$ is recurrent for $d = 1, 2$ and transient for $d \geq 3$.*

Proof 1: the mathematician's random walk. The proof is easy if we define a random walk in $\mathbb{Z}^d$ as a sum of independent random walks $S_n(i)$ in $\mathbb{Z}$, i.e. if we set

$$W_n = (S_n(1), S_n(2), \ldots, S_n(d)) \in \mathbb{Z}^d.$$

Then we have $u_n = 0$ if n is odd, and

$$u_{2n} = P(W_{2n} = 0) = P(S_{2n}(1) = 0)^d \sim \left(\frac{1}{\sqrt{\pi n}}\right)^d \asymp \frac{1}{n^{d/2}}.$$

Since $\sum n^{-d/2} < \infty$ iff $d \geq 3$, the preceding result completes the proof. ∎

Proof 2: the physicist's random walk. Alternatively one can define a random walk in $\mathbb{Z}^d$ so it only changes one coordinate at a time; then each step is in one of $2d$ possible directions, with equal probability. For $d = 1$ and $d = 2$ this is essentially the same as the walk above, so the analysis is the same.

We now prove transience for $d \geq 3$. For concreteness, assume $d = 3$. Then we have

$$u_{2n} = 6^{-2n} \sum \frac{(2n)!}{i!i!j!j!k!k!},$$

where the sum is over all partitions $i + j + k = n$. We can rewrite this as

$$u_{2n} = 2^{-2n} \binom{2n}{n} \sum \left(3^{-n} \frac{n!}{i!j!k!}\right)^2.$$

The terms which are squared come from the trinomial distribution, so their sum is no bigger than the largest term, which occurs when i, j and k are all about $n/3$. In other words, we have

$$u_{2n} = O\left(\frac{1}{\sqrt{n}} \frac{n!}{3^n((n/3)!)^3}\right).$$

Now by Stirling's formula, we have $n! \asymp \sqrt{n}(n/e)^n$ and hence $(3^n(n/3)!)^3 \asymp n^{3/2}(n/e)^n$. This gives $u_{2n} = O(1/n^{3/2})$, so $\sum u_{2n} < \infty$. ∎

Coda: The paradox of bad luck. Here are two paradoxes showing the persistence of bad luck (cf. Feller Volume 2, Ch. I.), also based on divergent series.

1. Suppose parking tickets are given out at random. At the end of the year, how many neighbors will I need to consult to find one who got more tickets than me? All things being equal, in a group of n I have a $1/n$ chance of having the most tickets. Since $\sum 1/n = \infty$, on average it will take forever to find a neighbor with worse luck. (The expected waiting time for first success is $\sum q_n$, where q_n is the probability of failure in the first n trials.)

2. Suppose you are in a traffic jam on 2-lane highway. It is impossible to change lanes. Every now and then, the right or left lane advances by one car. Each lane has the same chance of advancing.

   The red car next to you moves up one. How long do you expect it will take for you to catch up?

   This is the same as the waiting time for a first return to the origin in a random walk: it is infinite.

   The red car's driver hopes to eventually get two lengths ahead of you. However, his expected waiting time for this event is also infinite. And nevertheless, the gap between the two of you eventually gets arbitrarily large.

   To clarify the paradox — note that there is a big difference between saying the expected waiting time is infinite, and that you should expect to wait an infinite amount of time. In fact, you are certain to eventually pass the red car, and get as large a lead as you might wish.

   Keep in mind that any event which has the slightest positive chance of never happening has an expected waiting time of infinity. A single immortal vampire makes the average lifetime infinite.

Volume II

I The Exponential and the Uniform Density

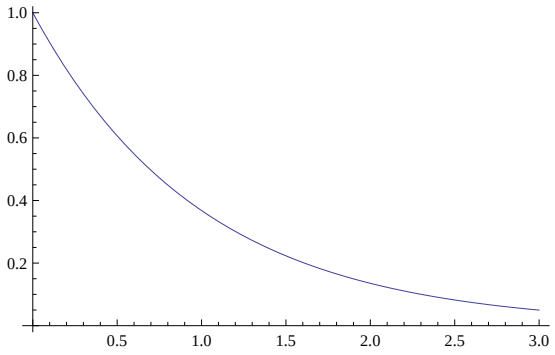

Figure 7. The exponential density with $E(X) = 1$.

Real-valued random variables. We now turn to the general study of *continuous* random variables, $X : S \to \mathbb{R}$. In this theory it is traditional to shift the focus from S itself to the (cumulative) distribution function

$$F(a) = P(X < a).$$

Note that $F(t)$ is a function which increases from 0 to 1 as t goes from $-\infty$ to ∞.

Often X will have a well-defined *probability density* $f(t)$, satisfying

$$P(a < X < b) = \int_a^b f(x)\,dx = F(b) - F(a).$$

In this case $f(x) = F'(x)$. We also have $f(x) \geq 0$ and $\int f(x)\,dx = 1$. Intuitively, if we repeatedly sample X, the probability that the result lies in the interval $[x, x + \Delta x]$ is approximately $f(x)\Delta x$.

Examples.

1. The *uniform density* on $[c, d]$ is given by $f(x) = 1/(c-d)$ on $[c, d]$ and $f(x) = 0$ elsewhere. For $[a, b] \subset [c, d]$ we have

$$P(a < X < b) = \frac{b-a}{c-d}.$$

The distribution function $F(t)$ increases linearly from $F(c) = 0$ to $F(d) = 1$. Clearly $E(X) = (c + d)/2$.

2. For $\alpha > 0$, the *exponential density* is given by

$$f(x) = \alpha \exp(-\alpha x)$$

on $[0, \infty)$. We have $F(t) = 1 - \exp(-\alpha t)$ for $t \geq 0$.

3. A single Bernoulli trial X has the distribution function $F(t) = 0$ for $t \leq 0$, $F(t) = q$ in $(0, 1]$, and $F(t) = 1$ for $t > 1$. Its ordinary derivative does not exist, but it is conventional to write

$$F'(t) = q\delta_0 + p\delta_1.$$

4. In the same way a discrete random variable with probability distribution p_k has

$$F(t) = \sum_{k<t} p_k \quad \text{and} \quad f(t) = \sum_k p_k \delta_k.$$

5. The *normal distribution* is given $f(x) = \mathfrak{n}(x) = (2\pi)^{-1/2} \exp(-x^2/2)$.

6. The *arcsine law* is given by the density $f(x) = (\pi x(1 - x))^{-1/2}$, with density $F(x) = (2/\pi) \arcsin(\sqrt{x})$ on $[0, 1]$.

What are the events? Here is a technical problem we are skirting. Suppose X is chosen uniformly at random in the interval $[0, 1]$. Then for 'any' $A \subset [0, 1]$ we have

$$P(X \in A) = |A|.$$

But what is $|A|$? For an interval this is clear, but what about for a more general subset of $[0, 1]$?

It turns out to be a technical challenge to identify exactly the sets for which the event $X \in A$ has a well-defined probability. A full resolution requires the theory of Lebesgue measure. Luckily we can skirt this issue for now! (As a warmup, one might ask: what is the probability that X is rational?)

From the discrete to the continuous. Continuous random variables emerge frequently as limits of discrete random variables. To make this precise, let us say (the distribution of) a random variable X_n *converges* to (that of) X if, for each $a \in \mathbb{R}$, we have

$$P(X < a) = \lim P(X_n < a).$$

(Note: it is important here that we use strict inequality at a.) Equivalently, their distribution functions satisfy $F_n(t) \to F(t)$ for all $t \in \mathbb{R}$. We can revisit the examples above from this perspective.

1. Let I_n be chosen at random among the numbers $(1, \ldots, n)$, each with equal probability $1/n$, and let $X_n I_n/n$. Then X_n converges to the uniformly distributed random variable X with density $f(x) = 1$ on $[0, 1]$.

2. Let F_n be the epoch of first success in a sequence of Bernoulli trials with $p = \alpha/n$. Recall that

$$P(F_n = k) = q^{k-1}p = (1 - \alpha/n)^k(\alpha/n) \approx \alpha \exp(-\alpha(k/n))(1/n).$$

   Then $X_n = F_n/n$ converges to the random variable X with the exponential density $f(x) = \alpha \exp(-\alpha x)$. Thus X can be regarded as the lifetime of a particle that has probability $\alpha \Delta x$ of decaying in time Δx.

3. Let S_n be a sequence of Bernoulli trials, and let $S_n^* = (S_n - E(S_n))/\sigma(S_n)$. Then S_n^* converges to the random variable with normal distribution $f(x) = (2\pi)^{-1/2} \exp(-x^2/2)$.

4. Let P_n be the number of steps a random walk of length n spends on the positive axis. Then P_n/n converges to a random variable X with the arcsine distribution $f(x) = (\pi x(1 - x))^{-1/2}$. (See section §III).

Expectations. Let X have density $f(x)$. Then we have

$$E(X) = \int x f(x)\, dx,$$

and more generally

$$E(g(X)) = \int g(x)\, f(x)\, dx.$$

In particular, $\mathrm{Var}(X)$ and $\sigma(X)$ can be defined as before.

We note that if $X \geq 0$, then

$$E(X) = \int_0^\infty x f(x)\, dx = \int_0^\infty 1 - F(x)\, dx.$$

Examples.

1. A uniformly distributed random variable on $[0,1]$ has expectations $E(X) = 1/2$ and, more generally, $E(X^n) = 1/(n+1)$. (These numbers are called the *moments* of X.) Thus $\mathrm{Var}(X) = E(X^2) - E(X)^2 = 1/12$, and $\sigma(X) = 1/2\sqrt{3} \approx 0.289$.

2. A random variable with the exponential density $f(x) = \alpha^{-1}\exp(-\alpha x)$ and distribution $F(x) = 1 - \exp(-\alpha x)$ satisfies

$$E(X) = \int_0^\infty 1 - F(x)\,dx = \int_0^\infty \exp(-\alpha x)\,dx = 1/\alpha.$$

   This is a continuous version of the fact that the expected waiting time for the first success in a sequence of Bernoulli trials is $1/p$. We also have $E(X^n) = n!/\alpha^n$ and hence $\mathrm{Var}(X) = \alpha^{-2}$, $\sigma(X) = 1/\alpha$.

3. The *Cauchy distribution* is given by

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

   By integrating we find $F(x) = (\pi/2 + \tan^{-1}(x))/\pi$. If we draw a line in $\mathbb{R}^2$ through $(0,1)$ at a random angle, then the location $(X,0)$ where it meets the x-axis defines a Cauchy distributed random variable.

   This gives an important example of a random variable $|X|$ with *infinite expectation*. (Notice that $xf(x) \sim 1/x$ for large x, and $\int_1^\infty dx/x = \infty$.)

Continued fractions. Every real number x can be expressed uniquely as a continued fraction

$$x = [a_0, a_1, \cdots] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}$$

with $a_i \in \mathbb{Z}$ and $a_i \geq 1$ for $i \geq 1$. The integers a_i are called the *partial quotients*.

If $x \in [0,1]$ is randomly chosen with respect to a uniform distribution, then we can regard the integers a_i as random variables. It turns out they are distributed according to the law

$$p_k = P(a = k) = \log_2\left(\frac{(k+1)^2}{k(k+2)}\right) \asymp \frac{1}{k^2}.$$

For example, $p_1 = \log_2(4/3) \approx 0.415$, $p_2 = \log_2(9/8) \approx 0.16$, etc. In particular, the expected value of the partial quotients is infinite!

Note that large values of a_i correspond to good rational approximations to x. For example, the continued fraction of π is given by

$$\pi = [3, 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, \ldots]$$

We have $\pi \approx [3, 7] = 3 + 1/7 = 22/7$, but even better $\pi \approx [3, 7, 15, 1] = 355/113 \approx 3.1415929$; this last is quite good because the partial quotient 292 is so large.

The continued fraction of π appears to have the same statistics as a random point, but it has not even been proved that the a_i's for π are unbounded.

Independent variables and convolution. We say X and Y are independent if for all a, b, c, d we have

$$P(a < X < b \text{ and } c < Y < d) = P(a < X < b) \cdot P(c < Y < d).$$

The definition for more than two random variables is analogous.

If X_1 and X_2 are independent with densities $f_1(x)$ and $f_2(x)$, then the density of $X = X_1 + X_2$ is given by the *convolution*

$$f(a) = \int f_1(x) f_2(a - x)\, dx = (f_1 * f_2)(a).$$

This follows from the fact that its distribution function satisfies

$$F(a) = P(X < a) = \int_{x_1 + x_2 < a} f_1(x_1) f_2(x_2)\, dx_1\, dx_2 = \int F_1(x) f_2(a - x)\, dx.$$

In both cases the result is symmetric in X_1 and X_2.

One should compare $f_1 * f_2$ to the *product of generating functions* in the case of discrete random variables.

Uniform and exponential examples. If X_1 and X_2 have a uniform distribution, then $X = X_1 + X_2$ has distribution

$$f(x) = \min(x, 2 - x), \quad x \in [0, 2].$$

Thus $F(t) = P(X < t) = t^2/2$ for $t \in [0, 1]$, and $1 - t^2/2$ for $t \in [1, 2]$. Note that $F(t)$ is simply the area of the intersection of the unit square $[0, 1] \times [0, 1]$ with the region $x_1 + x_2 \leq t$.

If X_1 and X_2 have the same exponential density $\alpha \exp(-\alpha s)$, then $X = X_1 + X_2$ has the density

$$f(s) = \int_0^s \alpha^2 \exp(-\alpha t) \exp(\alpha t - \alpha s)\, dt = \alpha^2 s \exp(-\alpha s).$$

Notice that $f(0) = 0$ every thought $f_1(0) = f_2(0) = 1$.

Ratios of variables. It also straightforward to compute the distribution of the ratio $X = X_1/X_2$ of two non-negative, independent random variables with densities $f_1(x)$ and $f_2(x)$. Indeed, since

$$G(t) = P(X < t) = P(X_1 < tX_2) = \int_0^\infty f_2(x) F_1(tx)\, dx,$$

the density of X is given by

$$g(t) = G'(t) = \int_0^\infty x f_2(x) f_1(tx)\, dx.$$

If X and $1/X$ have the same distribution, then we also have

$$g(t) = t^{-2} g(1/t).$$

Examples. (i) If X_1 and X_2 are chosen uniformly in $[0, 1]$, then for $t \leq 1$ we have $g(t) = 1/2$, and for $t > 1$ we have $g(t) = 1/(2t^2)$. (ii) If X_1 and X_2 have an exponential distribution (with any α), then

$$g(t) = \int_0^\infty \alpha^2 x \exp(-\alpha(1 + t)x)\, dx = (1 + t)^{-2},$$

and $1 - G(t) = 1/(1 + t)$. Thus when two people stand in the same line, half the time one has to wait 3 or more times longer than the other (since $P(X_1/X_2 > 3) = 1 - G(3) = 1/4$.

Moreover, in both examples, $E(X_1/X_2)$ is infinite. This is a common phenomenon; it arises just because $E(1/X_2)$ is infinite. Indeed, if the density $f(x)$ has a positive limit $f(0)$ as $x \to 0+$, then $E(1/X) = \infty$. To see this, note that the distribution function of $Y = 1/X$ is given by

$$P(Y < t) = P(X > 1/t) = 1 - F(1/t),$$

so the density of Y is given by

$$g(t) = f(1/t)/t^2,$$

and if $f(1/t) \to c \neq 0$ as $t \to \infty$, then

$$E(Y) = \int tg(t)dt \approx \int_1^\infty ct/t^2\, dt = \infty.$$

If X_1 is not identically zero, then $P(X_1 > a) > 0$ for some a, and thus $E(X_1/X_2) \geq P(X_1 > a)E(a/X_2) = \infty$.

More on exponentially distributed random variables. We now turn to a more detailed discussion of a random variable X with the exponential distribution,

$$f(x) = \alpha \exp(-\alpha x), \quad F(x) = 1 - \exp(-\alpha x);$$

note that $F'(x) = f(x)$. Here are some of the ways it arises naturally.

1. The function $L(s) = P(X > s) = 1 - F(s)$ satisfies the important relation:

   $$L(s+t) = L(s)L(t)$$

   for all $s, t > 0$, and is indeed characterized by this relation. If we think of X as the length of a lifetime, this means the probability of surviving for time $s+t$ is the probability of first surviving for time s and then surviving for an additional time t. The essential feature is that the risk of death is always the same, and the threats at different times are independent.

   Note that the expected lifetime is $E(X) = 1/\alpha$.

2. The probability that a thread of length t can sustain a fixed load satisfies $P(t+s) = P(t)P(s)$. So the length X at which the thread snaps is an exponentially distributed random variable.

3. The exponential distribution arises as the limit of the waiting time for first success in a sequence of Bernoulli trials with probability $p \to 0$, with $n \to \infty$ trials in each unit time interval, and with $np \to \alpha$. Note that α is the expected number of successes in a unit time interval. In the survival model this is the expected number of fatalities, which explains why the expected lifetime is $1/\alpha$.

   In the discrete case, the epoch of first success is exponentially distributed, with

   $$p_k = q^{k-1}p.$$

If we let $x = k/n$, then $\Delta x = 1/n = p/\alpha$, and we have

$$p_k \sim (1-p)^k p = (1-\alpha/n)^{nx} \alpha \Delta x \to \alpha e^{-\alpha x} \Delta x$$

as $n \to \infty$.

4. The Poisson process with density λ produces a random set $A \subset [0, \infty)$ with $E(|A \cap [a,b]) = \lambda |b-a|)$. We have seen that

$$P(|A \cap I| = k) = \exp(-\lambda |I|) \frac{\lambda^k |I|^k}{k!}.$$

This implies the location $X \geq 0$ of the smallest point in A is exponentially distributed: we have

$$P(X > t) = P(A \cap [0,t] = \emptyset) = \exp(-\lambda t).$$

Since the Poisson process is a limit of Bernoulli trials, it is no surprise that the location of its smallest point is a limit of the waiting time for the first success.

5. The *gaps* between consecutive points in the Poisson process have the *same* distribution as the smallest point; they are also exponentially distributed, with the expected gap size $1/\lambda$.

6. Fire a rifle in a thin forest, and let X be the distance the bullet travels before hitting a tree. Then X is exponentially distributed (to a good approximation).

Indeed, we can model the trees in the forest by a collection of disks of radius r in the plane whose centers are chosen using a Poisson process of density λ. (Some disks may overlap, but this is unlikely if the forest is thin, i.e. if λr^2 is small.)

If our bullet travels distance $\geq t$ then there are no centers within distance r of a given line of length t. The area of this region being $2rt$, we find

$$P(X > t) = \exp(-2\lambda r t)$$

by the definition of the Poisson process.

7. By the same token, the distance X between a tree and its nearest neighbor satisfies

$$P(X > t) = \exp(-\lambda \pi t^2),$$

since the event $X > t$ requires a region of area πt^2 to be free of trees. (Here our idealized trees have $r = 0$.)

Simulation of a disorganized bus system. How can one simulate a bus system with arrivals coming on average once every 10 minutes according to a Poisson process?

Suppose we have a way to generate a uniformly distributed random number Z in $[0,1]$, and we wish to model the random variable X. This can be done by simply setting $X = F^{-1}(Z)$, where F is the distribution function of X. In particular, an exponentially distributed random variable can be generated by setting

$$X = \frac{|\log Z|}{\alpha}.$$

We now simply choose independent waiting times $X_1, X_2, \ldots$ with $\alpha = 1/10$, and let the n bus arrive at time $T_n = X_1 + \cdots + X_n$. See Figure 8.

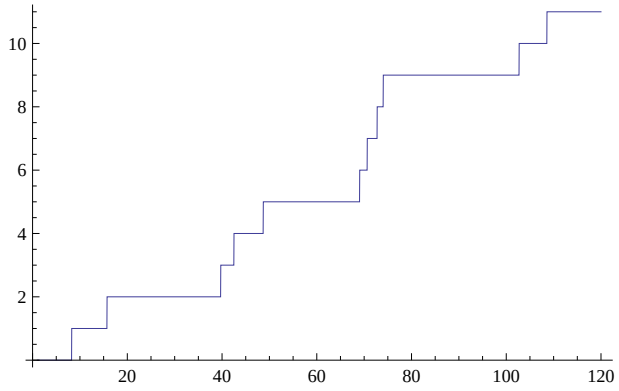

Figure 8. Bus arrivals, averaging one every 10 minutes.

The waiting time paradox. We can now revisit the waiting time paradox. Because of the lack of memory in the Poisson process, the expected waiting time till the next bus is 10 minutes. But we can also argue that our arrival is chosen at random in one of the gaps between buses. These gaps average 10 minutes, so our average waiting time should be 5 minutes.

Here is another way to phrase the paradox. The average waiting time for the next bus is 10 minutes. By the same reasoning, the average time since the previous bus arrived is 10 minutes. Thus the expected size of the gap between buses is 20 minutes, not 10 minutes.

Here is a resolution of the paradox. If $X_1, X_2, X_3, \ldots$ are the gaps between consecutive buses, then indeed $E(X_i) = 10$. That is, if we choose a gap at random, then its expected length is 10.

But if we choose a *time* at random, then we are more likely to land in a large gap than a small gap. That is, the probability of landing in a gap of size x should be proportional to the *product* of the length of the gap and the probability that a gap of that size occurs. So if we sample by random times, the gap size Y is distributed according to the density

$$g(y) = \alpha^2 y \exp(-\alpha y).$$

The constant α^2 is determined by the fact that $\int g(y)\,dy = 1$.

Notice that g is exactly the convolution $f * f$ for $f(x) = \alpha \exp(-\alpha x)$ computed above. In other words, Y is distributed as the sum $X_1 + X_2$ of two exponential waiting times with $E(X_i) = 10$. In particular $E(Y) = 20$. Its behavior is consistent with our reasoning that the length of time between buses is the sum of the time to the next bus and the time from the preceding bus.

Thus the paradox arises from ambiguity in the notion of a randomly chosen gap.

Completion of a task. Suppose an order for a new passport must pass consecutively through n independent levels of administration, each taking on average time $1/\alpha$ to complete their review. What is the distribution of the total time $S_n = X_1 + \cdots + X_n$ required to get the passport back?

We may assume X_i has the distribution $f(x) = \alpha \exp(-\alpha x)$, in which case we need only compute the n-fold convolution $f_n(x) = (f * f * \cdots * f)(x)$. This can be done inductively, and we find

$$f_n(x) = \alpha \frac{(\alpha x)^{n-1}}{(n-1)!} \exp(-\alpha x).$$

The case $n = 2$ was considered above. We have $E(S_n) = n/\alpha$ and $\mathrm{Var}(S_n) = n/\alpha^2$. By the central limit theorem, S_n^* limits to a normal distribution; equivalently, when suitably rescaled, the graph of the function $x^n e^{-x}$ approaches a bell curve.

Order statistics: the exponential post office. The zip code 27182 is served by a post office with n windows, each staffed by an exponential, independent worker who serves on average α customers per hour. This means the time X_i to service a customer at the ith window has the density $f(x) = \alpha \exp(-\alpha x)$.

Suppose when you arrive at the post office, each window is occupied by exactly one customer. No more customers are expected for the day.

1. How long will it take for a free window to open up?

The answer is $X_{(1)} = \min(X_1, \ldots, X_n)$. This variable has the distribution $f_1(x) = n\alpha \exp(-n\alpha x)$. To see this, recall that X_i is essentially a waiting time for a first success. Thus $X_{(1)}$ is also a waiting time for first success, but with trials proceeding n times faster.

For a proof based on calculation, just note that

$$P(X_{(1)} > t) = \prod_1^n P(X_i > t) = \exp(-n\alpha t).$$

Note: the Markov property (lack of memory) for the variables X_i means the amount of time a customer has already spent at the window has no effect on how long it will take for him to finish.

2. What are the chances I will be the last person to leave the post office?

   The answer is $1/n$. Once a window opens up, all the customers are treated equally, again by the Markov property.

3. Suppose I simply want to wait until all the customers leave, so I can rob the post office (take all their stamps). How long should I expect to wait?

   Let $X_{(i)}$ be the time it takes for a total of i customers to be serviced. We have already seen that $X_{(1)}$ has an exponential distribution with constant $n\alpha$. By the Markov property, $X_{(k)} - X_{(k-1)}$ has an exponential distribution with constant $(n-k)\alpha$. Hence

$$E(X_{(n)}) = \frac{1}{\alpha}\left(\frac{1}{n} + \frac{1}{n-1} + \cdots \frac{1}{1}\right) \approx \frac{\log n}{\alpha}.$$

   Note that the expected time required to service n customers at 1 window is n/α. Serving n customers with n windows is less efficient, because many windows are idle while the last few customers are being served.

4. What is the distribution of the time $X_{(k)}$ required for k of the n customers to be serviced? The event $X_{(k)} < t$ means that k or more of the events $(X_i < t)$ occur. Hence

$$F_k(t) = P(X_{(k)} < t) = \sum_{j=k}^{n} \binom{n}{j} (1 - \exp(-\alpha t))^j \exp(-\alpha t)^{n-j}.$$

By differentiating, one finds the density of $X_{(k)}$ is given by:

$$f_k(t) = n\binom{n-1}{k-1}(1-\exp(-\alpha t))^{k-1}\exp(-\alpha t)\exp(-\alpha t)^{n-k}.$$

This formula can be easily seen directly. If $X_{(k)}$ is in the interval $[t, t+\Delta t]$, this means $(k-1)$ customers have already finished, a kth customer finished in this interval, and the remaining $(n-k)$ customers are taking longer. This partition of n into 3 sets can be made in $n\binom{n-1}{k-1}$ ways. For each partition, the probability that the chosen $(k-1)$ customers have left is given by the first term in the product above, and the probability that the remaining $(n-k)$ customers are still waiting is given by the last term. The middle term, times Δt, gives the probability that the one remaining customer leaves during the time interval $[t, t+\Delta t]$.

Radioactive decay. It is common to measure the rate of decay of an unstable particle by its *half life* T. This is the *median* decay time, i.e. it is the solution to $\exp(-\alpha T) = 1/2$, namely $T = \log 2/\alpha$. (The *expected* time to decay is longer, namely $1/\alpha$.)

Thus the expected time for a sample of n atoms to decay completely is given by

$$E(X_{(n)}) = \frac{\log n}{\alpha} = T\log_2 n$$

This is intuitively clear — if the half the sample decays in time T, then the number of atoms left after time kT is $2^{-k}n$.

Example: Iodine 131 has a half-life of 8.02 days. (It mostly undergoes beta decay, in which a neutron emits an electron to become a proton, resulting in stable xenon). The mass of a single atom is about 131 u, where $u = 1.66 \times 10^{-24} g$. Thus a gram of I-131 contains about $n = 4.6 \times 10^{21}$ atoms; we have $\log_2 n \approx 72$, and hence it takes about 577 days to decay completely.

Of course a very small sample of I-131 poses little health risk. The same cannot be said for a small number of bacteria or infested plants. Recall the crop-duster analysis of Ch. IV.

The two envelopes problem.

The two envelopes problem is a puzzle or paradox within the subjectivistic interpretation of probability theory; more specifically within subjectivistic Bayesian decision theory. This is still an

open problem among the subjectivists as no consensus has been reached yet.

—Wikipedia, 4/2011.

Here is a variant of the Monty Hall paradox. Two envelopes hold 10 and 20 dollars respectively. You are presented with one at random. Would you like to switch to the other?

Of course it makes no difference. Your expected payoff is 15 in either case.

Now suppose you are told that one envelope holds twice as much as the other. You open the first envelope and find X dollars. If you switch, you will get either $X/2$ or $2X$ dollars. Since the average of these is $5X/4$, of course you should switch (?).

One way to clarify the paradox is this. There are indeed two possibilities; the envelope you hold contains C dollars, and the other contains $D = 2C$ or $D = C/2$. But more precisely, $D = 2C$ when $C = 10$ and $D = C/2$ when $C = 20$. It is a mistake to assume that D/C is independent of C!

Now let's make the problem more precise and devise a reasonable strategy. Let us suppose the envelopes hold A and $B = 2A$ dollars, where A is a random variable with finite expectation. Let us also suppose you know the distribution of A, and the value C in the first envelope.

In this case, one can sometimes give a good reason for switching. For example, if A is always odd, then you should switch whenever C is odd.

To handle the general case, we suppose A is chosen at random on $\mathbb{R}$ with respect to the density $f(x)$, that K is a Bernoulli trial with values 1 or 2 with equal probability, and that A and K are independent. Then $C = AK$.

Suppose we switch envelopes when $C \in E$. Then we gain A if $K = 1$ (in which case $A = C \in E$), but we lose A if $K = 2$ (in which case $A = C/2 \in E/2$). Thus our expected gain is

$$g = \frac{1}{2}\left(\int_E xf(x)\,dx - \int_{E/2} xf(x)\right)\,dx.$$

Now if $E = [s, s + \Delta s]$, then

$$2g \approx sf(s)\Delta s - (s/2)f(s/2)\Delta s/2 > 0$$

iff $f(s) \geq f(s/2)/4$. So we should switch if

$$C \in E = \{s \,:\, 4f(s) > f(s/2)\}.$$

Example: for the uniform distribution on $[0,1]$, we should switch if $C \in [0,1]$, but not switch if $C \in [1,2]$. (Of course in the latter case we know we hold the envelope containing $2A$.)

For an exponential distribution $f(s) = \alpha \exp(-\alpha s)$, we should switch when

$$C < (4 \log 2)/\alpha = (4 \log 2) E(A) \approx 2.772 E(A).$$

This is an example of *optimal decision–making* guided by probability theory.[1]

Optimal stopping times. Here is another classical optimization problem. In this year n movies will be released. Your goal is to pick the best movie of the year the moment it comes out (to seal your reputation as a reviewer).

The outcome is considered a success *only* if you correctly pick the best movie. You get no credit for picking second largest.

What strategy should you pursue to maximize the probability of success? And what is your probability of success, for large n?

Let us assume the movies have already been made, and the ith movie that is released has quality X_i (an objective amount), but otherwise nothing is known except the total number that will be released.

It turns out that any optimal strategy must have the following form:

> *Reject the first s movies, then select the first movie that is better than any you have seen so far.* mbut record that highest quality

As an example, let $s = n/2$. Then if $X_{(2)} = X_i$ for $i \in [1, s]$ while the maximum $X_{(1)} = j$ occurs for $j \in [s+1, n]$, then the strategy will success. These events are independent and each has probability $1/2$, so $s = n/2$ has a success rate for $1/4$.

To analyze the success rate for general s, suppose the maximum occurs at index $k \in [1, n]$. All these indices are equally likely, and we certain fail if $k \leq s$. What about for $k > s$? To succeed then, we need that:

> *The maximum of $(X_1, \ldots, X_{k-1})$ occurs at an index $i \leq s$.*

For if it occurs at an index $i \in [s+1, k-1]$, we will acccept this 'false maximum' as our final result. But otherwise we will wait until X_k.

Now the probability of the event above is just $s/(k-1)$, again because each of the indices in $[1, k-1]$ are equally likely. So the probability of success

[1] Here is another envelope problem: suppose you do not know the distribution of the amount placed in the envelopes. Devise a strategy for switching which does better, on average, than never switching at all.

is:

$$p(s) = \frac{1}{n} \sum_{k=s+1}^{n} \frac{s}{k-1} \approx \frac{s}{n} \int_s^n \frac{dx}{dx} = \frac{s}{n} \log \frac{n}{s}.$$

The function $f(x) = x \log(1/x)$ has $f'(x) = \log(1/x) - 1 = 0$ at $x = 1/e$, where it assumes the value $f(1/e) = 1/e$.

> *Thus the optimal strategy is to reject the first* $\approx n/e \approx 0.368n$ *numbers, and then accept the next record breaker.*

The success rate is similarly $1/e \approx 36.8\%$. Note that there are two types of failure: finding no new record, and accepting a new record that is still not optimal. The first type only happens when the maximum value of X_i occurs in the interval $[1, s]$, so its probability is also $s/n \approx 1/e$. A false maximum is chosen the rest of the time, i.e. in $1 - 2/e = 26.4\%$ of the cases.

More on uniformly distributed random variables. When one speaks of a point X chosen at random in $[0, 1]$, this usually means chosen at random with respect to the *uniform* density. That is, $P(X \in A) = |A|$ for any interval $A \subset [0, 1]$.

A sequence of independent, uniformly distributed random variables

$$(X_1, \ldots, X_n) \in [0, 1]^n$$

is the same as a random point in a cube. Such a sequence cuts $[0, 1]$ into $(n + 1)$ intervals, $[0, X_{(1)}]$, $[X_{(1)}, X_{(2)}]$, $\cdots [X_{(n)}, 1]$. As before, the *order statistics* $(X_{(1)}, \ldots, X_{(n)})$ are obtained by putting the sequence $(X_1, \ldots, X_n)$ into increasing order.

Theorem I.1 *The lengths* $(L_1, \ldots, L_{n+1})$ *of the complementary intervals* $X_{(1)}$, $X_{(2)} - X_{(1)}$, …, $1 - X_{(n)}$ *have a common distribution, satisfying:*

$$1 - F_n(t) = P(L_i > t) = (1 - t)^n \quad \textit{and} \quad f_n(t) = n(1 - t)^{n-1}.$$

The expected length of a complementary interval is $1/(n + 1)$.

Proof. Imagine instead that we choose $(n + 1)$ independent points, uniformly distributed on the unit circle $S^1 = \mathbb{R}/\mathbb{Z}$. It is then clear that all the gaps have the same distribution.

To compute the value of $1 - F(t)$, we just note that the condition $L_1 = X_{(1)} = \min(X_1, \ldots, X_n) > t$ means that each X_i lies in the interval $(t, 1]$, and the volume of the subcube $(t, 1)^n$ in $[0, 1]^n$ is $(1 - t)^n$. Since $\sum L_i = 1$, we have $E(L_i) = 1/(n + 1)$. ∎

Corollary I.2 *The expected value of $X_{(k)}$ is $k/(n+1)$.*

Proof. $E(X_{(k)}) = E\left(\sum_1^k L_i\right)$. ∎

Theorem I.3 *The order statistics $X_{(k)}$ for a uniform distribution of n points in $[0,1]$ have densities*

$$f_k(t) = n\binom{n-1}{k-1}t^{k-1}(1-t)^{n-k}.$$

Proof. This is similar to the exponential case: to get $X_{(k)}$ in $[t, t+\Delta t]$ we need to have $k-1$ points with $X_i < t$; if their indices are fixed, then the probability of this event in t^{k-1}. We also need $n-k$ points in $[t,1]$; this contributes a factor of $(1-t)^{n-k}$. The one remaining point must land in $[t, t+\Delta]$; the probability of this event is Δt. Finally the number of ways of partitioning n into these three sets is $n\binom{n-1}{k-1}$. ∎

Note that $f_k(t)/n$ is exactly the probability that among $n-1$ Bernoulli trials of probability t, we have $k-1$ successes.

Correlation. Unlike the order statistics for the exponential distribution, the lengths $L_1, \ldots, L_{n+1}$ are correlated, as can be seen immediately from the fact that $\mathrm{Var}(\sum L_i) = 0$. For example, when $n=1$, we have $L_2 = 1 - L_1$, and hence $\rho(L_1, L_2) = -1$.

The uniform post office. In the uniform post office, the service time required at the ith window is a uniformly distributed random variable in $[0,1]$. As before, each window has exactly one customer when I arrive.

1. What is the expected wait for a free window to open up? This is $E(X_{(1)}) = E(L) = 1/(n+1)$.

2. How long will it take, on average, for all customers n to be serviced? Answer: $E(X_{(n)}) = n/(n+1)$.

3. Suppose we have $n = 2k-1$ windows. How is the departure time of the middle customer, $X_{(k)}$, distributed?

   Answer: it behaves like a normally distributed random variable with $E(X) = 1/2$ and $\sigma(X) = 1/(2\sqrt{n})$. Indeed, if $t = 1/2 + s$, then the density $f_k(t)$ is a constant multiple of

$$\begin{aligned}(1+2s)^k(1-2s)^k &= (1-4s^2)^k = \left(1 - \frac{4ks^2}{k}\right)^k \\ &\sim \exp(-4ks^2) = \exp(-s^2/2\sigma^2),\end{aligned}$$

where $\sigma^2 = 1/8k \approx 1/4n$.

Roulette, Benford's law and the uniform distribution. The number of total turns T of a roulette will be fairly smoothly distributed over a large interval $[0, M]$. The outcome of the game is given by the random variable $X = T \bmod 1$. It is easy to believe, and not hard to prove, that X will be close to uniformly distributed if M is large and the distribution is smooth enough. For example, if the density $g(t)$ of T is increasing for $t < a$ and decreasing for $t > a$, and $f(a) = m$, then the density $f(x)$ of X satisfies $|f(x)-1| \le m$. This is because $f(x) = \sum_n g(x+n)$ is a good approximation to $\int g(t)\,dt = 1$.

By the same token, for any random variable T with a large range (many orders of magnitude) and rather smooth distribution, $X = \log_{10} T \bmod 1$ will be approximately uniformly distributed. But the value of the first digit N of T depends only on X; this gives rise to *Benford's law*:

$$P(N = k) = \frac{\log(k+1) - \log k}{\log 10}.$$

Thus $P(N = 1) = \log_{10} 2 \approx 30.1\%$, while $P(N = 9) = 1 - \log_{10} 9 \approx 4.6\%$.

Random directions; other densities. Suppose X has the density $f(x)$ and $X = \phi(Y)$, where ϕ is $1-1$. Then the density $g(y)$ of Y is given by

$$g(y) = f(\phi(y))|\phi'(y)|.$$

This is intuitively clear: if Y is $[y, y+\Delta y]$, then X is in $[\phi(y), \phi(y)+\phi'(y)\Delta y]$ and the probability of the latter event is $f(\phi(y))|\phi'(y)|\Delta y$. It can also be seen using distribution functions. If ϕ is increasing, then

$$G(t) = P(Y < t) = P(X < \phi(t)) = F(\phi(t)),$$

so $g'(t) = f(\phi(t))\phi'(t)$.

In particular, if X is uniformly distributed, then $g(y) = \phi'(y)$. For example, if X is chosen uniform in $[-1/2, 1/2]$, and $Y = \tan(\pi X)$, then $X = (1/\pi)\tan^{-1}(Y)$ and hence $g(y) = 1/(\pi(1+y^2))$. This is the source of the Cauchy distribution.

Theorem I.4 *The tangent of a random angle is distributed according to the Cauchy density* $f(x) = 1/(\pi(1+x^2))$.

Projections.

Theorem I.5 *The length of the projection to a line of a random vector on the unit circle $S^1 \subset \mathbb{R}^2$ has density*

$$f_{21}(x) = \frac{2}{\pi\sqrt{1-x^2}}$$

in $[0,1]$. *Its expected value is* $2/\pi > 1/2$.

Proof. The length L is given by $\sin(\theta)$ for a random angle $\theta \in [0,\pi]$. Thus

$$E(L) = \frac{1}{\pi}\int_0^\pi \sin\theta\, d\theta = \frac{2}{\pi}.$$

To compute the density, we just note that

$$dL = -\cos\theta\, d\theta,$$

and so

$$\frac{|d\theta|}{\pi} = \frac{|dL|}{\pi\cos\theta} = \frac{|dL|}{\pi\sqrt{1-L^2}}.$$

■

Theorem I.6 *The length L_1 of the projection to a line of a random vector on the unit sphere $S^2 \subset \mathbb{R}^3$ is uniformly distributed in* $[0,1]$*; thus* $E(L_1) = 1/2$. *The length L_2 of its projection to a plane has density*

$$f_{32}(x) = \frac{x}{\sqrt{1-x^2}}$$

and $E(L_2) = \pi/4$.

Proof. We will use the egg-coloring theorem from calculus: the area of zone $a < z < b$ on S^2, with $a, b \in [-1,1]$, is proportional to $|b-a|$. The first result is then immediate, by projecting to the z-coordinate. For the second, we use the fact that $L_2^2 + L_1^2 = 1$ for orthogonal planes. Thus

$$2L_2 dL_2 + 2L_1 dL_1 = 0,$$

which gives

$$|dL_1| = \frac{L_2|dL_2|}{L_1} = \frac{L_2|dL_2|}{\sqrt{1-L_2^2}}.$$

As an exercise in spherical coordinates, with latitude $\theta \in [0,2\pi]$ and $\phi \in [0,\pi]$ the distance from the north pole, we have $dA = \sin\phi\, d\phi\, d\theta$ and $\int dA = 4\pi$. Projecting to the equatorial plane gives $L_2 = \sin\phi$, and so

$$E(L_2) = \frac{2\pi}{4\pi}\int_0^\pi \sin^2\phi\, d\phi = \frac{\pi}{4}.$$

■

Small lengths. Note that a small length is much more likely for a projection from S^1 to a line than from S^2 to a plane. Indeed, $f_{21}(0) = 2/\pi$ while $f_{32}(0) = 0$.

Example: cosmic rays. What is the expected distance D a cosmic ray hitting a particle detector has spent traveling through the interior of earth?

The ray travels along a random line L through our observatory N (say at the north pole). The length S of the intersection of L with the earth is the same as the length of the projection of the north-south axis NS to L. Thus it is a uniformly distributed random variable with $E(S) = r$, where r is the radius of the earth.

Now half the time the particle comes from the sky, and half the time from the earth. Thus $E(D) = r/2$. The function D itself is equal to 0 with probability 1/2, and is uniformly distributed in $[0, 2r]$ with probability 1/2. This is a natural example of a mixture of continuous and discrete densities.

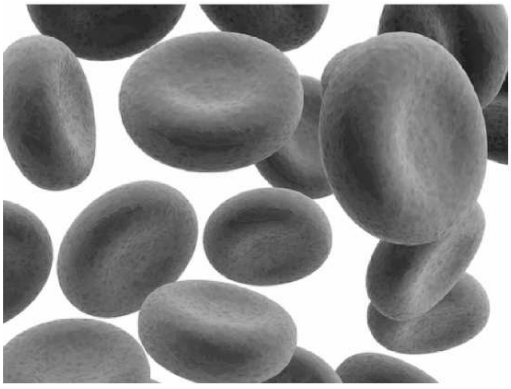

Figure 9. Red blood cells.

Example: red cells. A red blood cell (erythrocyte) is essentially a flat round disk of radius R. When viewed through a microscope, one sees it in profile as an ellipse with major and minor radii r and R. How is the eccentricity r/R distributed?

We assume the unit normal to the plane of the cell is uniformly distributed on S^2. Then the length of the projection of the normal to the line of sight L is uniformly distributed in $[0, 1]$. But the angle between the normal and the line of sight is the same as the angle between the plane of the cell and the plane of sight. From this we see that r/R is also uniformly distributed.

Note: it is sometimes assumed that the angle θ between the normal line to the cell and the line of sight is uniformly distributed in $[0, \pi]$. But this is wrong! And if it were true, one would see many more nearly round-looking corpuscles.

The Buffon needle problem. A needle of unit length is dropped onto a floor covered by wooden slats of unit width. What is the probability that the needle lies athwart two strips?

Think of the slats as the lines $y \in \mathbb{Z}$ in $\mathbb{R}^2$. The needle is described by the location $Y \in [0, 1]$ of its center, within the slat it lies upon, and the angle $\theta \in [0, \pi]$ it makes with respect to the slats. We assume both are uniformly distributed. Its projection to the y-axis has length $L = \sin\theta$. The needle crosses a slat if $Y + L/2 > 1$ or if $Y - L/2 < 0$. These two possiblities are exclusive, so the probability that one or the other happens is

$$p = 2P(Y < L/2) = \frac{2}{\pi}\int_0^\pi \frac{1}{2}\sin\theta\, d\theta = \frac{2}{\pi}.$$

Thus a spilled box of needles can be used to calculate the ratio between the diameter and circumference of a circle. In 1901, Mario Lazzarini did exactly that, tossing a needle 3408 times to obtain an estimate with an error of less than 10^{-6}. (He used needles that were 5/6th the width of the slats of wood.) This implausible level of accuracy is not unrelated to the strange number of tosses: it is critical that $3408 = 213n$ for $n = 16$, since $213 = 355 \cdot (5/3)$, and this leads to the very good approximation $5\pi/3 \approx 113/213$ if the number of flips comes out the right multiple of 113.

Correction. Chap I.13 (29): Ignore the hint: the distribution of radii is irrelevant.

II Special Densities. Randomization

In this section we discuss Gaussian random variables and the Cauchy distribution, and relate their stability properties to their characteristic functions (Fourier transforms) $E(\exp(itX))$. The latter material is drawn from Chapter XV.

Gaussian random variables. A random variable Y has a *Gaussian distribution* if it is of the form $Y = aX + b$, where X is the standard Gaussian with $E(X) = 0$ and $\sigma(X) = 1$. The density of X is given by

$$\mathfrak{n}(x) = \frac{1}{2\pi}e^{-x^2/2},$$

so the constants a and b are uniquely determined by $m = E(Y)$ and $\sigma = \sigma(Y)$; and in fact the density of Y is given by

$$f(y) = \frac{1}{\sigma}\mathfrak{n}\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

(In general, if X has distribution $n(x)$ and $X = aY + b$, then Y has distribution $an(ay + b)$.)

Theorem II.1 *The sum $Y = X_1 + X_2$ of two or more independent Gaussian random variables is again Gaussian. Indeed, Y is the unique Gaussian with $E(Y) = E(X_1) + E(X_2)$ and $\sigma(Y)^2 = \sigma(X_1)^2 + \sigma(X_2)^2$.*

Proof. It suffices to show that the density $f(y)$ has the form $A\exp Q(y)$, where Q is a quadratic polynomial (complete the square). To this end we note that

$$f(y) = \int A_1 A_2 \exp(Q_1(x) + Q_2(y-x))\,dx$$

for suitable quadratic polynomials $Q_i(x)$. Their sum has total degree 2 in (x, y), so it can be written in the form

$$Q_1(x) + Q_2(y-x) = Q(y) + (Ax + By + C)^2.$$

(That is, we can choose A, B and C to absorb the x^2, xy and x terms into the second factor). Since $\int \exp((Ax+By+C)^2)\,dx$ is a constant, independent of y, we conclude that $f(y)$ is a constant multiple of $e^{Q(y)}$ and hence Gaussian. ∎

Remark. As a special case, if $S_n = X_1 + \cdots X_n$, then S_n/n has the same distribution as X_1. This is clear from the central limit theorem for Bernoulli trials: we can think of each X_i as the limit of repeated Bernoulli trials, and S_n just repeats them more.

Characteristic functions. The *characteristic function* of a random variable X is defined by for $t \in \mathbb{R}$ by

$$\phi(t) = E(\exp(itX)).$$

If $X = \sum p_k \delta_k$ only takes integer values, then we have

$$\phi(t) = \sum p_k \exp(itk).$$

Thus ϕ is obtained from the generating function $\sum p_k s^k$ by substituting $s = \exp(it)$; it records the values of the generating function on the unit circle in the complex plane. In this case $\phi(t + 2\pi) = \phi(t)$.

On the other hand, if X has density $f(x)$, then

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx.$$

Thus $\phi(t)$ is essentially the *Fourier transform* of the function $f(x)$.

Let us note some formal properties of the characteristic function.

1. We have $\phi(0) = 1$ and $|\phi(t)| \leq 1$.
2. If $Y = aX + b$, then $\phi_Y(t) = \exp(itb)\phi_X(at)$.
3. We have $\phi'(0) = E(iX)$, $\phi''(0) = -E(X^2)$, and $\phi^{(n)}(0) = i^n E(X^n)$, assuming the moments $E(X^n)$ exist.

Another critical property is:

Theorem II.2 *If X and Y have the same characteristic function, then they have the same distribution.*

But the most crucial connection with probability theory comes from:

Theorem II.3 *The characteristic function of a sum $X = X_1 + X_2$ of two independent random variables is given by $\phi(t) = \phi_1(t)\phi_2(t)$.*

Proof. We have $E(\exp(it(X_1 + X_2))) = E(\exp(itX_1))E(\exp(itX_2))$. ∎

We can now revisit the stability properties of the Gaussian.

Theorem II.4 *The characteristic function of a Gaussian random variable X is another Gaussian, of the form*

$$\phi(t) = \exp(ibt)\exp(-a^2t^2/2). \tag{II.1}$$

We have $b = 0$ if $E(X) = 0$.

Proof. By the properties above, it suffices to treat the case of the standard Gaussian with $E(X) = 0$ and $E(X^2) = 1$. Then by completing the square, we have

$$\begin{aligned} \phi(t) &= \frac{1}{\sqrt{2\pi}} \int \exp(-(x^2 - 2it)/2)\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int \exp(-(x - it)^2/2)\exp(-t^2/2)\, dx = \exp(-t^2/2). \end{aligned}$$

∎

Since the product of two Gaussian function as in (II.1) is again of the same form, we find:

Corollary II.5 *The sum $Y = X_1 + \cdots X_n$ of two or more Gaussian random variables has a Gaussian distribution.*

The exact distribution is determined by $E(Y) = \sum E(X_i)$ and $\mathrm{Var}(Y) = \sum \mathrm{Var}(X_i)$.

Law of large numbers. Let $S_n = X_1 + \cdots X_n$ be a sum of independent random variables all with the same characteristic function ϕ. Then the characteristic function of S_n is given by $\phi_n(t) = \phi(t)^n$. Hence the characteristic function of S_n/n is given by

$$\mu_n(t) = \phi(t/n)^n.$$

For example, if X_1 is Gaussian then $\mu_n(t) = \exp(-t^2/2n) \to 1$ as $n \to \infty$. This is a reflection of the fact that $S_n/n \to 0$, and it holds more generally if $E(X) = 0$ and $E(X^2)$ is finite.

Cauchy random variables. Recall that the standard Cauchy random variable X with density $f(x) = 1/(\pi(1 + x^2))$ has $E(|X|) = \infty$. Let us see how this behavior is reflected in its characteristic function.

Theorem II.6 *The characteristic function of the standard Cauchy random variable is given by*

$$\phi(t) = \exp(-|t|).$$

Proof. This computation is best done by complex analysis; for $t > 0$, the function $\exp(itz)$ is rapidly decreasing as $\mathrm{Im}\, z \to \infty$, and $1 + z^2$ vanishes at $z = i$; thus by the residue theorem, we have

$$\phi(t) = \frac{1}{\pi} \int_{\mathbb{R}} \frac{\exp(itz\, dz)}{1 + z^2} = \frac{2\pi i}{\pi} \mathrm{Res}_{z=i} \frac{\exp(itz)}{1 + z^2} = 2i \frac{\exp(-t)}{2i} = \exp(-t).$$

For $t < 0$ we get $\phi(t) = \exp(+t)$. ∎

Note that $\phi'(0)$ does not exist. However the right and left handed derivatives exist and sum to zero. This reflects the fact that the Cauchy principal value of the expectation is zero, i.e. $\lim_{N\to\infty} \int_{-N}^{N} x\, dx/(1 + x^2) = 0$.

Failure of averaging. What happens if we take repeated Cauchy observations and average them?

Theorem II.7 *The average $S_n/n = (X_1 + \cdots + X_n)$ of a sequence of independent, identically distributed Cauchy random variables has the same distribution as each one.*

Proof. We have $\mu_n(t) = \phi(t/n)^n = \exp(-|t|/n)^n = \exp(-|t|)$. ∎

This gives a useful stability property, as well as a strong example of failure of averaging to converge.

The central limit theorem. We can now indicate a conceptually simple proof of the central theorem, although some technical details need to be verified to make it fully rigorous.

Suppose $X_1, \ldots, X_n$ are independent random variables with a common distribution, and $E(X_i) = 0$ and $E(X_i^2) = 1$. Let $\phi(t)$ be the corresponding characteristic function, and let $\phi_n(t)$ be the characteristic function of

$$S_n^* = \frac{X_1 + \cdots X_n}{\sqrt{n}}.$$

We then have

$$\phi_n(t) = \phi(t/\sqrt{n})^n.$$

Now our assumptions on mean and variance imply

$$\phi(t) = 1 - t^2/2 + o(t^2).$$

Thus

$$\phi_n(t) \sim \left(1 - \frac{t^2}{2n}\right)^n \to \exp(-t^2/2)$$

as $n \to \infty$, and the right hand side is the characteristic function of the standard Gaussian. This shows that S_n^* converges to the standard Gaussian distribution, at least on the level of characteristic functions.