AN INTRODUCTION TO PROBABILITY
AND RANDOM PROCESSES

by

Kenneth Baclawski

and

Gian-Carlo Rota

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

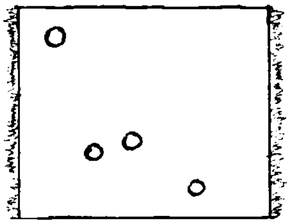
Probability is one of the great achievements of this century. Like geometry, it is a way of looking at nature. There are many ways of approaching natural problems, many points of view. The geometrical point of view has been with us for thousands of years. The probabilistic point of view is another way of focusing on problems that has been successful in many instances. The purpose of this course is to learn to think probabilistically. Unfortunately the only way to learn to think probabilistically is to learn the theorems of probability. Only later,as one has mastered the theorems, does the probabilistic point of view begin to emerge while the specific theorems fade in one's memory: much as the grin on the Cheshire cat.

We begin by giving a bird's-eye view of probability by examining some of the great unsolved problems of probability theory. It's only by seeing what the unsolved problems are that one gets a feeling for a field. Don't expect to be able to understand at this point everything about the problems we are about to give. They are difficult and are meant to be just a hint of things to come.

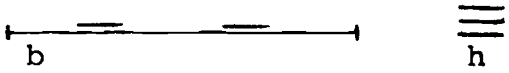
Pennies on a carpet. We have a rectangular carpet and an indefinite supply of perfect pennies. What is the probability that if we drop the pennies on the carpet at random no two of them will overlap? This problem is one of the most important problems of statistical mechanics. If we could answer it we would know, for example, why water boils at 100°C, on the basis of purely atomic computations. *Nothing* is known about this problem.

On the other hand, the one-dimensional version of this problem can be solved. We shall, in fact, solve it several different ways. The problem here is to drop n needles of length h on a stick of length b at random. The probability that no two needles overlap is:

$$\begin{cases} \left(\dfrac{b-nh}{b-h}\right)^{n} & \text{if } b \geq nh \\ 0 & \text{if } b<nh \end{cases}$$



Pennies on a carpet



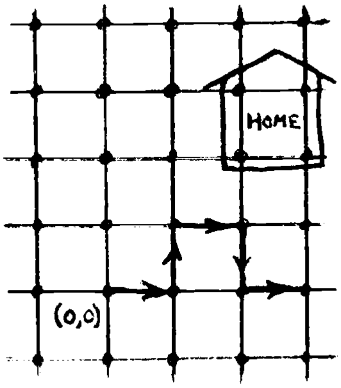
Needles on a stick

The striking difference between the difficulty of a problem in two dimensions and that of the corresponding problem in one dimension is called the "dimensional barrier". It is an illustration of a common problem of physicists: problems in low dimensions are considerably easier to solve than their "real world" counterparts.

The only technique that we can presently apply to this problem is the "method of ignorance" or the "Monte Carlo method": **namely** simulate the problem on a computer, and see what happens. Usually a few iterations will give a remarkably accurate answer, when the number of coins is small.

Random walk. We consider the grid on the plane with integral corners. A drunkard starts at the origin and walks in one of the four directions with equal probability 1/4 to the next corner. He then repeats this process at the next corner.

It is already an interesting mathematical question to set up this problem so that one can answer such questions as for example how long it will take the drunkard to get home. The answer is not a number but rather a probability distribution; that is, there is a certain probability that he will get home in 1 step, 2 steps, 3 steps, etc.



a random walk

So this is a typical case for which we ask a question, and we get an answer that is not a number but rather a string of numbers each with a suitable probability. We call this "answering a question probabilistically" or "using probabilistic reasoning."

One can completely answer the above question, given the position and shape of the drunkard's home. This connects with a branch of physics called potential theory.

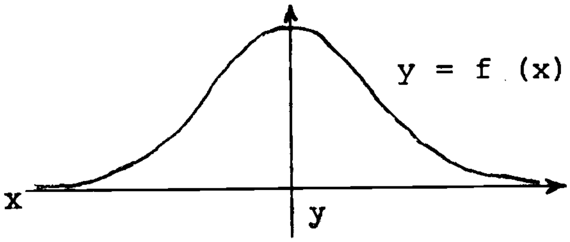
A question that has never been answered is to find the probability that after n steps the drunkard has *never* retraced his steps. We call such a random walk a *self-avoiding random walk*. This is related to the problem of polymer growth in chemistry. Of course it appears that this is a very special, stereotyped problem, but it turns out that if we can solve this stereotyped problem, we can solve all the others by suitable coordinate changes. This will be the case also in problems we shall subsequently encounter.

Cluster analysis. Suppose that we have a collection of dots arrayed in the plane or in space, much as stars in the sky. The individual points obey no specific physical law, but the whole ensemble does. The problem is to invent the possible physical laws that such ensembles of dots can satisfy.

For example how can one describe that a pattern of dots obeys certain clustering structures? Physically it is sometimes quite obvious: we just look in the sky. But we want a purely numerical, quantitative description. This is the theory of stochastic point processes.

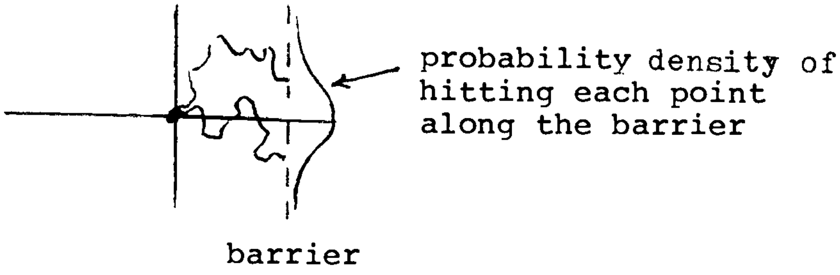
Of course this problem is closely related to the problem of pattern recognition.

Brownian Motion. We must first mention a functjon called the normal density function: $f\ (x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. This function occurs so often in nature as well as in probability that one is tempted to call it the most important function there is. It looks like this:



this is the famous "bell-curve".

Now a realistic model of the path of a drunkard is one that wanders in a continuous path starting at the origin. How does one assign probabilities to paths and what does it mean to follow a path at random? This was done by Norbert Wiener who showed that if there is a straight line barrier in the plane and if we consider the question of where the drunkard first hits the barrier we get precisely the normal density function.



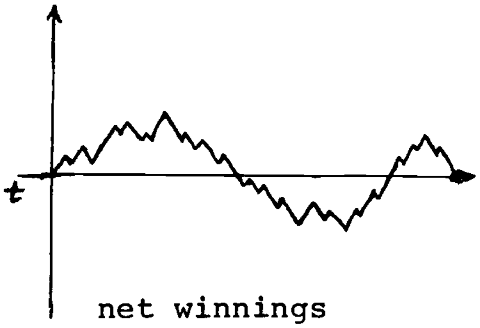
This fact enables us to compute, just as with the case of discrete random walks, the probability distribution for when the drunkard arrives home in terms of the position and shape of his home. This is the problem of Brownian motion. Unlike the others, this problem has been completely solved.

Contagion or Percolation. Imagine that we have an orchard with evenly spaced trees and that at some time some trees become infected. Suppose that there is a certain probability that a given infected tree infects one or more neighboring trees before the given tree dies.

One of two things can happen: either the infection stays among small clusters of infected trees and eventually dies out or the whole orchard is wiped out. One can show that if the probability of one infected tree infecting another is p there is a critical probability p_c such that if $p<p_c$ the disease will die out but that if $p>p_c$ the disease will spread forever. How does one compute p_c?

Noise. We consider a signal sent from a radio transmitter to a receiver but which is perturbed by noise along the way. The problem of filtering out the noise is a very important one for electrical engineers. The whole theory of noise filtering consists of computations involving the normal density function.

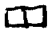
Coin Tossing. The detailed structure of the fluctuations occurring in the tossing of a fair coin are counter-intuitive. We imagine a game for which at each toss of a fair coin we win \$1 if it comes up heads and we lose \$1 if it comes up tails. If we graph our net winnings in time we see that it can cross the time axis if we switch from a net gain to a net loss or vice versa. If after a period of time we find that we have a net gain of zero, what is the most probable number of times we crossed over the axis along the way? The answer is that the most probable case is no times at all!



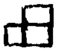
In effect one can interpret this as saying that during a long betting session the most probable occurrence is to have either a winning streak or a losing streak. Frequent changes from one to the other are actually unlikely.

Cell Growth. How does living tissue grow? We consider a stereotyped case. We start with a little square and then imagine that with some probability the square produces a new square on one of its four sides. The growth proceeds on the boundary by a simple model. What is the pattern that such growth will produce? What is the probability that the tissue will enclose an island?

(a)



(b)



(c)

* * *

The problems described above are just a sampling of the many interesting unsolved problems of probability. Perhaps you will be the one to solve one of them...

Chapter I Sets, Events and Probability

Suppose that we toss a coin any number of times and that we list the information of whether we got heads or tails on the successive tosses:

H	T	H	H	T	T	T	.	.	.
1	2	3	4	5	6	7	.	.	.

The act of tossing this coin over and over again as we have done is an example of an *experiment*, and the sequence of H's and T's that we have listed is called its *experimental outcome*. We now ask what it means, in the context of our experiment, to say that we got a head on the fourth toss. We call this an *event*. While it is intuitively obvious what an event represents, we want to find a precise meaning for this concept. One way to do this which is both obvious and subtle is to *identify* an event with the set of all ways that the event can occur. For example, the event "the fourth toss is a head" is the same as the set of all sequences of H's and T's whose fourth entry is H. At first it appears that we have said little, but in fact we have made a conceptual advance. We have made the intuitive notion of an event into a concrete notion: *events are a certain kind of set*.

However a warning is appropriate here. An event is not the same concept as that of an experimental outcome. An outcome consists of the total information about the experiment after it has been performed. Thus while an event may be easy to describe, the set to which it corresponds consists of a great many possible experimental outcomes, each being quite complex. In order to distinguish the concept of an event from the concept of an experimental outcome we will employ an artificial term for the latter. We will call it a _sample point_. Now a sample point will seldom look like an actual point in the geometric sense. We use the word "point" to suggest the "indivisibility" of one given experimental outcome, in contrast to an event which is made up of a great many possible outcomes. The term "sample" is suggestive of the random nature of our experiment, where one particular sample point is only one of many possible outcomes.

We will begin with a review of the theory of sets, with which we assume some familiarity. We will then extend the concept of a set by allowing elements to occur more than just once. We call such an entity a _multiset_. By one more conceptual step, the notion of a probability measure emerges as an abstraction derived from the multiset concept. Along the way we will repreatedly return to our coin-tossing experiment. We do this not only because it is a good example but also because we wish to emphasize that probability deals with very special kinds of sets.

1. The Algebra of Sets

In probability we always work within a _context_, which we define by specifying the set of all possible experimental outcomes or equivalently all possible sample points. We call this set the sample space, typically denoted Ω. The term "sample space" does not help one to visualize Ω any more than "sample point" is suggestive of an experimental outcome. But this is the term that has become standard. Think of Ω as the "context" or "universe of discourse". It does not, however, in itself define our experiment. Quite a bit more will be required to do this. One such requirement is that we must specify which subsets of Ω are to be the events of our experiment. In general not every subset will be an event. The choice of subsets which are to be the events will depend on the phenomena to be studied.

We will specify the events of our experiment by specifying certain very simple events which we will call the "elementary events", which we then combine to form more complicated events. The ways we combine events to form other events are called the Boolean or logical operations. The most important of these are the following:

union $A\cup B$ is the set of elements either in A or in B (or both).

intersection $A\cap B$ is the set of elements both in A and in B.

complement $\bar{A}$ is the set of elements not in A.

Each of these has a natural interpretation in terms of events. Let A and B be two events.

$A \cup B$ is the event "either A or B (or both) occur"
$A \cap B$ is the event "both A and B occur"
$\bar{A}$ is the event "A does not occur"

Several other Boolean operations are defined in the exercises.

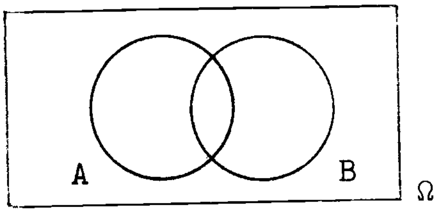
When two events A and B have the property that if A occurs then B does also (but not necessarily vice versa), we will say that A is a _subevent_ of B and will write $A \subseteq B$. In set-theoretic language one would say that A is a _subset_ of B or that B _contains_ A.

The three Boolean operations and the subevent relation satisfy a number of laws such as commutativity, associativity, distributivity and so on, which we will not discuss in detail, although some are considered in the exercises. For example the _DeMorgan laws_ are the following:

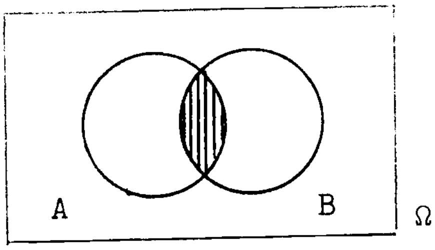
$$\overline{A \cap B} = \bar{A} \cup \bar{B}$$

$$\overline{A \cup B} = \bar{A} \cap \bar{B}.$$

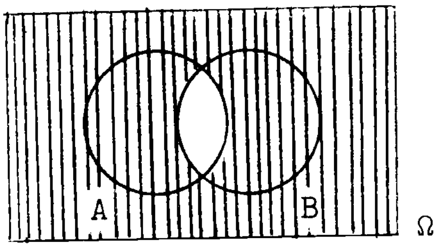
In terms of events, the first of these says that if it is not true that both A and B occur, then either A does not occur or B does not occur (or both), and conversely. One has a similar statement for the second law. Generally speaking, drawing a Venn diagram suffices to prove anything about these operations. For example, here is the Venn diagram proof of the first De Morgan law. First draw the two events A and B:



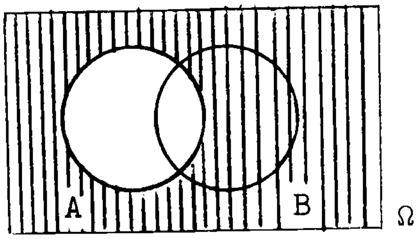
If we shade in the event $A \cap B$:

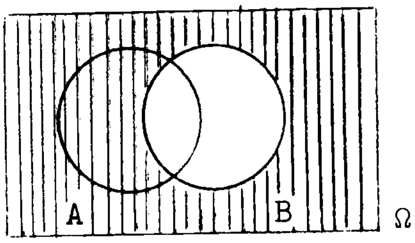


then the event $\overline{A \cap B}$ consists of the shaded portion of the following diagram:

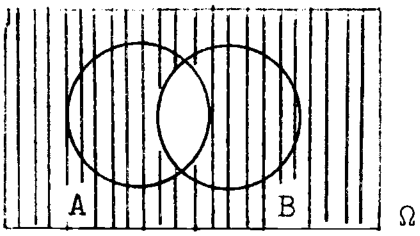


Next shade in the events $\bar{A}$ and $\bar{B}$, respectively:





Combining these gives us the event $\bar{A}\cup\bar{B}$:



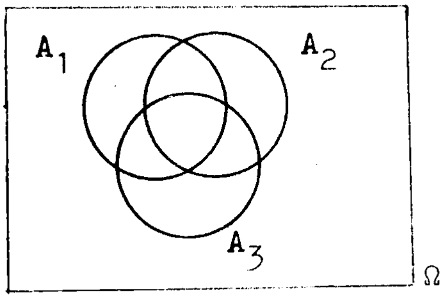
If we compare this with the event $\overline{A\cap B}$ we see that $\overline{A\cap B} = \bar{A}\cup\bar{B}$.

For more complicated expressions, involving many sets, and for which the Venn diagram would be extremely complex, it is very useful to know that there is a way we can simplify such an expression into an essentially unique expression. The idea is that every Boolean expression is a union of the smallest subevents obtainable by intersecting events occurring in the expression. To be more precise suppose that we have an expression involving the events $A_1, A_2, \ldots, A_n$ and unions, intersections and complements in any order nested as deeply as required. The simplified expression we obtain can be described in two steps.

Step 1. Write $A^{+1} = A$ and $A^{-1} = \bar{A}$. This is just a notational convenience; it has no metaphysical significance. The expressions

$$A_1^{i_1}\cap A_2^{i_2}\cap\cdots\cap A_n^{i_n} \quad ,$$

as $i_1, i_2, \cdots, i_n$ take on all possible choices of ± 1, are the smallest events obtainable from the events $A_1, A_2, \cdots, A_n$ using Boolean operations. We call these events the atoms defined by $A_1, A_2, \cdots, A_n$. Notice that in general there can be 2^n of them, but that in particular cases some of the atoms may be empty so there could be fewer than 2^n in all.



the events A_1, A_2, A_3 break up Ω into (at most) eight atoms.

Step 2. Any expression involving $A_1, A_2, \cdots, A_n$ and using Boolean operations can be written as a union of certain of the atoms. There are many procedures that can be used to determine which of the atoms are to be used. We leave it as an exercise to describe such a procedure. The resulting expression will be called the "atomic decomposition". By using Venn diagrams and atomic decompositions, any problem involving a finite number of events can be analyzed in a straightforward way. Unfortunately many of our problems will involve infinitely many events and for these we will later need some new ideas.

2. The Bernoulli Sample Space

We now return to the example that began this chapter: tossing a coin. A sample point for this experiment is an infinite sequence of ones and zeroes or equivalently of H's and T's. Just for variety we will also sometimes refer to a toss of heads as a "success" and tails as a "failure". Even if we are only concerned with a finite sequence of H's and T's, which is seemingly more realistic, it is nevertheless easier for computational reasons to imagine that we could go on tossing the coin forever. Moreover, we will find that certain seemingly rather ordinary events can only be expressed in such a context.

The set of all possible sequences of ones and zeroes is called the Bernoulli sample space Ω and the experimental process of which it forms the basis is called the Bernoulli process. For the moment we will be a little vague about what a "process" means, but we will make it precise later. The events of the Bernoulli sample space consist of certain subsets of Ω. To describe which subsets these are, we first describe some very simple events called elementary events. They are the events "the first toss comes up heads", "the second toss comes up heads", etc. We will write

H_n for the event "the n^{th} toss comes up heads". The complement of the event H_n will be devoted $T_n = \overline{H}_n$; it is the event "the n^{th} toss comes up tails." One must be careful here. It is obvious that the complementary event to "the n^{th} toss is heads" is "the n^{th} toss is tails." However, it is less obvious that as <u>sets</u> $\overline{H}_n = T_n$, since both H_n and T_n are infinite sets and it is not easy to imagine what they "look like". As a rule it is much easier to think in terms of the events themselves rather than in terms of their representations as sets.

We can now describe in general what it means for a subset of Ω to be an event of the Bernoulli smaple space: an event is any subset of Ω obtainable from the elementary events by using the operations of complement as well as of unions and intersections of infinite sequences of events. The fact that we allow infinite unions and intersections will take some getting used to. What we are saying is that we allow any statements about the Bernoulli process which may in principle be expressed in terms of tosses of heads and tails (elementary events) using the words "and" (intersection), "or" (union), "not" (complement) and "ever" (infinite sequences).

To illustrate this we consider the following example of a Bernoulli event: "a sequence of two successive H's occurs before a sequence of two T's ever occurs." We will call a specified finite sequence of H's and T's a <u>run</u>. So the event in question is "the run HH occurs before the run TT ever occurs". Write A for this event. The essence of the event A is that we will continue flipping that coin until either an HH or a TT occurs. When one of them happens we may then quit, or we may not;but it is nevertheless computationally easier to conceive of the experiment as having continued forever. Now it ought to be conceptually clear that it is possible to express A in terms of elementary Bernoulli events, but at first it may seem mysterious how to do it. The idea is to break apart A into simpler events which can each be expressed relatively easily in terms of elementary events. The whole art of probability is to make a judicious choice of a manner of breaking up the event being considered. In this case we break up the event A according to when the first run of HH occurs. Let A_n be the event "the run HH occurs first at the n^{th} toss and the run TT has not yet occurred." The event A is the (infinite) union of all the A_n's, and in turn each A_n can be expressed in terms of the elementary events as follows:

$$\begin{array}{ll}
A_2 = H_1 \cap H_2 & \text{HH ...} \\
A_3 = \overline{H}_1 \cap H_2 \cap H_3 & \text{THH...} \\
A_4 = H_1 \cap \overline{H}_2 \cap H_3 \cap H_4 & \text{HTHH...} \\
A_5 = \overline{H}_1 \cap H_2 \cap \overline{H}_3 \cap H_4 \cap H_5 & \text{THTHH...} \\
\qquad\qquad \text{etc.} &
\end{array}$$

Note that not only is A the union of the A_n's but also none of the A_n's overlap with any other. In other words no sample point of A has been counted twice. This property will be very important for probabilistic computations.

As an exercise one might try to calculate the expression, in terms of elementary events, of the event " a run of k heads occurs before a run of n tails occurs." Later we will develop tools for computing the probability of such an event quite easily, and this exercise will quickly convince one of the power of these tools.

3. The Algebra of Multisets

We now go back to our study of set theory. Our objective is to extend the concept of a set by allowing elements of sets to be repeated. This more general concept is called a multiset. To give an example, suppose that a committee of 10 members has an election to determine its chairperson. Of the votes that are cast, 7 are for candidate A, 2 for B and 1 for C. The set of votes is most easily expressed as a multiset consisting of 10 elements: 7 of type A, 2 of type B and 1 of type C. In set-builder notation we write this $\{a,a,a,a,a,a,a,b,b,c\}$. We can write this more economically as $\{a^7,b^2,c^1\}$, the exponents denoting the number of copies of the element that are in the multiset. Notice that a set is a special kind of multiset.

As with sets, we can combine multisets to form new multisets. In some ways these operations are more natural than the analogous ones for sets. The operations are *addition* and *multiplication*. In the exercises we describe one more operation: *subtraction*. Given two multisets M and N, their *sum* $M+N$ is obtained by combining all the elements of M and N, counting multiplicities. For example if a occurs three times in M and twice in N, then it occurs five times in $M+N$. The *product* MN of M and N is obtained by multiplying the multiplicities of elements occuring in both M and N. For example if a occurs three times in M and twice in N, then it occurs six times in MN. Here are some more examples:

$$\{a,a,a,b,b\} + \{a,b,b,b,c\} = \{a,a,a,a,b,b,b,b,b,c\}$$
$$\{a,a,a,b,b\} \cdot \{a,b,b,b,c\} = \{a,a,a,b,b,b,b,b,b\} ,$$
$$\{a,b\} + \{b,c\} = \{a,b,b,c\}$$
$$\{a,b\} \cdot \{b,c\} = \{b\}$$

or using exponent notation:

$$\{a^3,b^2\} + \{a^1,b^3,c^1\} = \{a^4,b^5,c^1\}$$
$$\{a^3,b^2\} \cdot \{a^1,b^3,c^1\} = \{a^3,b^6\} .$$
$$\{a^1,b^1\} + \{b^1,c^1\} = \{a^1,b^2,c^1\}$$
$$\{a^1,b^1\} \cdot \{b^1,c^1\} = \{b^1\}$$

When A and B are two sets, it now makes sense to speak of their sum A+B and their product AB. What do these mean in terms of sets? The product is easy to describe: it coincides precisely with the intersection $A\cap B$. For this reason it is quite common to write AB for the intersection of two events. On the other hand, the sum of two sets is not so easy to describe. In general A+B will not be a set even when both A and B are. The reason is that those elements occurring both in A and in B will necessarily occur twice in A+B. However if A and B are disjoint, that is when $A\cap B$ is empty, then A+B is a set and coincides with $A\cup B$. As this situation is quite important in probability, we will often write A+B to denote the union of A and B when A is disjoint from B, and we will then refer to A+B as the disjoint union of A and B.

4. The Concept of Probability

Consider once again the election multiset introduced in the last section: $\{a^7,b^2,c^1\}$. What percentage of the votes did each of the candidates receive? An easy calculation reveals that A received 70% of the total, B received 20% and C received only 10%. The process of converting "raw counts" into percentages loses some of the information of the original multiset, since the percentages do not reveal how many votes were cast. However, the percentages do contain all the information relevant to an election.

By taking percentages we are replacing the complete information of the number of votes cast for each candidate by the information of what proportion of votes were cast for each candidate relative to the total number of votes cast. The concept of probability is an abstraction derived from this situation. Namely, a probability measure on a set tells one the proportion or size of an element or a subset relative to the size of the set as a whole. We may intuitively think of a probability as an assignment of a non-negative real number to every element of the set in such a way that the sum of all such numbers is 1. The above multiset $\{a^7,b^2,c^1\}$ gives rise to a probability measure which will be denoted in the following manner. For every subset S of $\{a,b,c\}$, we write $P(S)$ for the proportion of the multiset $\{a^7,b^2,c^1\}$, which has elements from S. For example, $P(\{a\})$ is 0.7 because 70% of the elements of $\{a^7,b^2,c^1\}$ are a's. Similarly, $P(\{a,b\})$ is 0.9, $P(\{b,c\})$ is 0.3, $P(\{a,b,c\})$ is 1.0 and so on. We call P a probability measure. It is important to realize that P is defined not on elements but on subsets. We do this because we observed that events are subsets of the sample space, and we wish to express the concept of a probability directly in terms of events. As we have seen it is easier to think directly in terms of events rather than in terms of sets of outcomes. For this reason we henceforth decree that a probability measure P on a sample space Ω is a function which assigns a real number $P(A)$ to every event A of Ω such that

(1) $P(A) \geq 0$

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

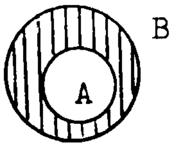
(3) If $A_1, A_2, \cdots$ is a sequence of disjoint events, then $P(A_1 + A_2 + \cdots) = P(A_1) + P(A_2) + \cdots$ or more compactly: $P(\sum_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

At first it may not be easy to see that these three conditions capture the concept of "proportion" we described above. The first two conditions however are easy to understand: we do not allow outcomes to occur a negative number of times, and the measure of Ω itself is 1 because it is the totality of all possible outcomes. It is the third condition that is the most difficult to justify. This condition is called *countable additivity*. When the sequence of events consists of just two events A and B, it is obvious. Let C be the union $A \cup B$. Since A and B are assumed to be disjoint, C is the same as $A + B$. Probabilistically this says that A and B are mutually exclusive alternatives for C: it occurs if and only if exactly one of A or B occurs. Clearly if this is so then the probability of C is "distributed" between A and B, i.e. $P(C) = P(A) + P(B)$. The extension of this rule to an infinite sequence of events is somewhat unintuitive, but one can get used to it when one sees concrete examples.

Properties of Probability Measures

We now show three important facts about probability measures. These facts relate the concept of probability to the Boolean concepts of subevent, union, intersection and complement.

Subevents. *If* A *is a subevent of* B, *then* $P(A) \leq P(B)$. Although this should be intuitively clear, we will *prove* it from the three conditions for P to be a probability. First observe that $A \subseteq B$ means that B is the disjoint union of A and $B \setminus A$, where $B \setminus A$ denotes $B \cap \bar{A}$. This should be clear from the Venn diagram, or just think of what it says: every element of B is either in A or it is not and these alternatives are mutually exclusive. Therefore condition (3) implies that



$B \setminus A$ is shaded

$$P(B) = P(A + (B \setminus A)) = P(A) + P(B \setminus A) \ .$$

By condition (1), $P(B \setminus A) \geq 0$. Therefore,

$$P(B) = P(A) + P(B \setminus A) \geq P(A) \ .$$

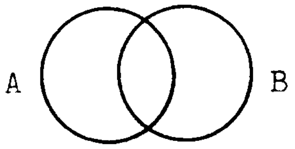
As a consequence we find that since every event A is a subevent of Ω ,

$$0 \leq P(A) \leq P(\Omega) = 1 \ .$$

This corresponds to our intuitive feeling that probability is a measure of likelihood, ranging from extremely unlikely (zero or near zero) to extremely likely (1 or close to 1).

Union and Intersection. *If A and B are two events, then* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

To prove this we first write $A \quad B$ as a disjoint union of atoms.



From the Venn diagram it is clear that

$$A \cup B = (A \cap B) + (A \setminus B) + (B \setminus A) \ .$$

Similarly, we can write A and B as (disjoint) unions of atoms:

$$A = (A \cap B) + (A \setminus B)$$

$$B = (A \cap B) + (B \setminus A) \ .$$

By condition (3),

$$P(A \cup B) = P(A \cap B) + P(A \setminus B) + P(B \setminus A)$$
$$P(A) = P(A \cap B) + P(A \setminus B)$$
$$P(B) = P(A \cap B) + P(B \setminus A).$$

Now solve for $P(A \setminus B)$ and $P(B \setminus A)$ in the last two expressions and substitute these into the first. This gives our formula. The usefulness of this formula is that it applies even when A and B are not disjoint.

Here is a concrete example. Suppose that we have two coins. Let A be the event that the first shows heads, and let $a = P(A)$ be the probability of this. Similarly let B be the event that the second shows heads, and let b be $P(B)$. What is the probability that when we toss both of them at least one shows heads? Clearly we want $P(A \cup B)$. By the above formula, we find that $P(A \cup B) = P(A) + P(B) - P(A \cap B) = a + b - P(A \cap B)$. However, we do not yet know how to compute $P(A \cap B)$ in terms of $P(A)$ and $P(B)$. We will return to this problem in the next section.

<u>Complement</u>. <u>If</u> A <u>is</u> <u>an</u> <u>event</u>, <u>then</u> $P(\bar{A}) = 1 - P(A)$. To see this simply note that Ω is the disjoint unior of A and $\bar{A}$. By conditions (2) and (3), we have $1 = P(\Omega) = P(A + \bar{A}) = P(A) + P(\bar{A})$. Thus we see that the probability for an event <u>not</u> to occur is "complementary" to the probability for its occurrance. For example, if the probability of getting heads when we toss a coin is p, then the probability of getting tails is $q = 1 - p$.

5. Independent Events

The notion of independence is an intuitive one derived from experience: two events are independent if they have no effect on one another. More precisely if we have two independent events A and B, then knowing A has occurred does not change the probability for B to occur and vice versa. When we have the notion of conditional probability we can make this statement completely rigorous. Nevertheless even with the terminology we have so far, the concept of independence is easy to express. We say two events A and B are independent when

$$P(A \cap B) = P(A)P(B)$$

If we use multiset notation, writing AB for $A \cap B$, then this rule is very suggestive: $P(AB) = P(A)P(B)$. It is important to realize that only independent events satisfy this rule just as only disjoint events satisfy additivity: $P(A+B) = P(A) + P(B)$.

Consider the case of coin tossing. The individual tosses of the coin are independent: the coin is the same coin after each toss and has no menory of having been tossed before. As a result, the probability of getting two heads in two tosses is the square of the probability of getting one head on one toss.

As an application consider the two-coin toss problem in the last section. Since we are tossing two different coins, it seems reasonable to expect A and B to be independent. Therefore

$P(A \cap B) = P(A)P(B) = ab.$ Thus

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= a + b - ab. \end{aligned}$$

We conclude that the probability for one of the coins to show heads is $a + b - ab$.

For any three events A,B and C, we say these events are <u>independent</u> when:

(1) any pair of the three are independent,

(2) $P(A \cap B \cap C) = P(A)P(B)P(C)$.

It is possible for three events to satisfy (1) but not (2). This is an important point that is easily missed. Consider again the two-coin toss problem above. Let C be the event that the two coins show different faces (one heads the other tails). Then A,B and C are pairwise independent; for example, knowing that the first coin shows heads tells one nothing about whether the other will be the same or different. However the three events are not independent: the occurrence of any two of them precludes the third from occurring.

Similarly given any number of events (even an infinite number), we say that they are <u>independent</u> when

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_2)P(A_2) \cdots P(A_n)$$

for <u>any</u> finite subcollection $A_1, A_2, \cdots, A_n$ of the events.

6. The Bernoulli Process

This is the process of tossing a biased coin. In a given toss we suppose that the probability is p for heads and q for tails, where $p+q = 1$. Generally speaking we will also be implicitly assuming that both p and q are nonzero, but other than this we shall make no restriction on what value p could have. We call p the bias of the coin. A fair coin is a special kind of biased coin; namely, one with bias $p = \frac{1}{2}$.

We want to assign a probability to every elementary event and show how this allows us to compute the probability of every other event. This done in two steps.

Step 1. $P(H_n) = p$ and $P(T_n) = P(\overline{H_n}) = q = 1-p$. This assignment is made irrespective of n. In effect we assume that we are using the same coin (or at least have the same bias) during each toss.

We have now defined the probability of the elementary events. But we still cannot determine the probability of an arbitrary event because noneof the three conditions determines what, for example, $P(H_1 \cap H_2)$ can be, although they limit the possible choices. This leads us to our second assumption:

Step 2. $P(H_1^{i_1} \cap H_2^{i_2} \cap \cdots \cap H_n^{i_n}) = P(H_1^{i_1})\, P(H_2^{i_2}) \cdots P(H_n^{i_n})$, where $i_1, i_2, \cdots, i_n$ take on all possible choices of ± 1.

Here we have drawn from the physical nature of the phenomenon of tossing a coin. The question of whether tosses of a real coin are independent is another question entirely: the question of the validity of the model as a means of describing the actual physical experiment. We will not consider this question until Chapter IV.

For any other event A of the Bernoulli process, the probability is calculated by expanding A in terms of the elementary events (possibly using countable unions) and by using conditions (1), (2) and (3). It is not always obvious what the best way to do this might be. There is an "art" to this part of the subject, and developing one's skill in this art is the whole idea of learning probability.

Let us return to the event: "a run of HH occurs before a run of TT." Recall that this event can be expanded as a disjoint union:

$$A = (H_1 \cap H_2) + (T_1 \cap H_2 \cap H_3) + (H_1 \cap T_2 \cap H_3 \cap H_4) + \cdots .$$

By countable additivity, we compute:

$$\begin{aligned}
P(A) &= P(H_1 \cap H_2) + P(T_1 \cap H_2 \cap H_3) + P(H_1 \cap T_2 \cap H_3 \cap H_4) + \cdots \\
&= P(H_1)P(H_2) + P(T_1)P(H_2)P(H_3) + P(H_1)P(T_2)P(H_3)P(H_4) + \cdots \\
&= p^2 + qp^2 + pqp^2 + qpqp^2 + \cdots \\
&= p^2 + pqp^2 + \cdots \\
&\quad + qp^2 + qpqp^2 + \cdots \\
&= p^2 \ (1 + pq + (pq)^2 + \cdots) \\
&\quad + qp^2 \ (1 + pq + (pq)^2 + \cdots) \\
&= p^2 \cdot \frac{1}{1 - pq} + qp \cdot \frac{1}{1 - pq} \\
&= \frac{p^2 + qp^2}{1 - pq} .
\end{aligned}$$

We are assuming here that we know how to sum a geometric series:

$$1 + r + r^2 + r^3 + \cdots = \frac{1}{1-r} \; , \quad \text{if} \quad |r| < 1 \; .$$

We can check our computation of $P(A)$ by a simple expedient: suppose that the coin is *fair*, i.e. $p = q = 1/2$. In this case $P(A) = 1/2$, for either HH occurs before TT or TT occurs before HH , and since the coin is fair either one of these is equally likely. And indeed setting $p = q = 1/2$ in our formula above shows that this is the case.

Probability measure: a function P on events such that

(1) for every event A, $P(A) \geq 0$

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

(3) if A_1, A_2, ... is a sequence of disjoint events, then $P(A_1+A_2+\cdots) = P(A_1)+P(A_2)+\cdots$

Properties of probability measures:

(1) if A and B are disjoint events, then $P(A+B)=P(A)+P(B)$

(2) if A and B are independent, then $P(AB)=P(A)P(B)$

(3) if A is a subevent of B, then $P(A) \leq P(B)$

(4) if A and B are any two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(5) if A is any event, then $P(\overline{A}) = 1-P(A)$

Probability Measures

7. Exercises for

Chapter I Sets, Events and Probabilty

The Algebra of Sets and Multisets

1. If A and B are sets, the stroke of A by B, written $A\setminus B$, stands for $A\cap\bar{B}$, i.e. those elements of A that are not in B. As an event $A\setminus B$ stands for "A occurs but B does not." Show that the operations of union, intersection and complement can all be expressed using only the stroke operation.

2. The symmetric difference of A and B, written $A \Delta B$, is defined by

$$A \Delta B = (A\setminus B)\cup(B\setminus A) .$$

As an event $A \Delta B$ means "either A occurs or B occurs but not both." For this reason this operation is also called the "exclusive or." Use a Venn diagram to illustrate this operation.

3. The set of elements where A implies B, denoted A/B, is the set

$$A/B = \bar{A}\cup B .$$

As an event A/B stands for "if A occurs then B does also." Use a Venn diagram to illustrate this operation.

4. Using Venn diagrams prove the following:

(a) $(A/B)\cap(B/C) \subseteq A/C$, i.e. if A implies B and B implies C, then A implies C.

(b) $(A/B)\cap(A/C) = A/(B\cap C)$, i.e. A implies B and A implies C if and only if A implies B and C.

(c) $(A/B)\cap(B/A) = \overline{A \Delta B}$.

5. Show that for any four sets A, B, C and D, the following is true: $(A \bigcup B) \setminus (C \bigcup D) \subseteq (A \setminus C) \bigcup (B \setminus D)$.

6. Prove that any Boolean (logical) expression in $A_1, A_2, \cdots, A_n$ is a union of atoms. In what sense is this union unique?

7. Let B be a multiset. We say that A is a <u>sub-multiset</u> of B if every element of A occurs at least as many times in B as it does in A. For example, {a, a, b} is a sub-multiset of {a, a, b, b, c} but not of {a, b, b,}. When A is a sub-multiset of B, it makes sense to speak of the <u>difference</u> of A and B, written B - A; namely define B - A to be the unique multiset such that A + (B - A) = B. For example, {a, a, b, b, c} — {a, a, b} = {b, c}. Suppose henceforth that A and B are sets. When does B - A coincide with $B \setminus A$? Prove that $A \bigcup B = A + B - AB$. Compare this with property (b) in section I.3.

The Bernoulli Sample Space

8. Give an explicit expression for the event "a run of three heads occurs before a run of two tails" in terms of elementary Bernoulli events. Suggest how this might be extended to "a run of k heads occurs before a run of n tails."

The Concept of Probability

9. In a certain town there are exactly 1000 families that have exactly three children. Records show that 11.9% have 3 boys, 36.9% have 2 boys, 38.1% have 2 girls and 13.1% have 3 girls.

Use a multiset to describe this situation. Give an interpretation in terms of probability. What is the probability that in one of the above families all the children have the same gender?

10. In a factory there are 100 workers. Of the total, 65 are male, 77 are married and 50 are both married and male. How many workers are female? What fraction of the female workers are married? Ask the same questions for male workers.

11. Express $P(A \cup B \cup C)$ in terms of $P(A)$, $P(B)$, $P(C)$, $P(A \cap B)$, $P(A \cap C)$, $P(B \cap C)$ and $P(A \cap B \cap C)$. Note the similarity of this expression with that of property (b) of section I.3.

12. Let D_1 be the event "exactly one of the events A, B and C occurs." Express $P(D_1)$ in terms of $P(A)$, $P(B)$, $P(C)$, $P(A \cap B)$, $P(A \cap C)$, $P(B \cap C)$ and $P(A \cap B \cap C)$.

13.* Condition (3) for a probability measure can be stated in several other ways. Prove that condition (3) implies each of (a) and (b) below and that either of these imply condition (3).

(a) If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ is an ascending sequence of events and if $A = A_1 \cup A_2 \cup A_3 \cup \cdots$, then

$$P(A) = \lim_{n \to \infty} P(A_n) .$$

(b) If $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$ is a descending sequence of events and if $A = A_1 \cap A_2 \cap A_3 \cap \cdots$, then

$$P(A) = \lim_{n \to \infty} P(A_n) .$$

Independent Events

14. If A and B form a pair of independent events, show that the pair $A, \bar{B}$, the pair $\bar{A}, B$ and the pair $\bar{A}, \bar{B}$ are each a pair of independent events.

15. In exercise 10, are the properties of being male and of being female independent? Is the property of being male independent of being married?

16. The probability of throwing a "6" with a single die is 1/6. If three dice are thrown independently, what is the probability that exactly one shows a "6"? Use exercise 12.

17. A student applies for two national scholarships. The probability that he is awarded the first is 1/2, while the probability for the second is only 1/4. But the probability that he gets both is 1/6. Are the events that he gets the scholarships independent of one another. Discuss what this means.

18. A baseball player has a 0.280 batting average. What is the probability that he gets exactly one hit in the next three times at bat? See exercise 16. To do this exercise one must assume that the player's times at bat are independent. Is this a reasonable assumption?

19. Three pigeons have been trained to peck at one of two buttons in response to a visual stimulus and do so correctly with probability p . Three pigeons are given the same stimulus. What is the probability that the majority peck at the correct stimulus? Suppose that one of the pigeons sustains an injury

and subsequently pecks at one or the other button with equal probability. Which is more likely to be the correct response, the button pecked by one of the normal pigeons or the button pecked by a majority of the three pigeons?

20. How many times must one roll a die in order to have a 99% chance of rolling a "6"? (Answer: 26 times.) If you rolled a die this many times and it never showed a "6", what would you think?

The Bernoulli Process

21. A dice game commonly played in gambling houses is Craps. In this game two dice are repeatedly rolled by a player, called the "shooter," until either a win or a loss occurs. It is theoretically possible for the play to continue for any given number of rolls before the shooter either wins or loses. Computing the probability of a win requires the full use of condition (3). Because of the complexity of this game, we will consider here a simplified version.

The probability of rolling a "4" using two (fair) dice is 1/12, and the probability of rolling a "7" is 1/6. What is the probability of rolling a "4" before rolling a "7"? This probability appears as part of a later calculation.

22. (Prendergast) Two technicians are discussing the relative merits of two rockets. One rocket has two engines, the other four. The engines used are all identical. To ensure success the engines are somewhat redundant: the rocket will achieve its

mission even if half the engines fail. The first technician argues that the four-engine rocket ought to be the better one.

The second technician then says, "Although I cannot reveal the failure probability of an engine because it is classified top secret, I can assure you that either rocket is as likely to succeed as the other."

The first technician replies, "Thank you. What you just told me allows me to compute the failure probability both for an engine and for a rocket."

Can you do this computation also?

23.* Let A be the event in the Bernoulli process that the coin forever alternates between heads and tails. This event consists of just two possible sample points.

$$A = \{ HTHTH\cdots, THTH\cdots \} .$$

Using exercise 13, prove that $P(A) = 0$. Does this mean that A is impossible? More generally if an event B in the Bernoulli process consists of only a finite number of sample points, then $P(B) = 0$.

Chapter II Finite Processes

Historically the theory of probability arose from the analysis of games of chance, usually based on the toss of a die or the choice of a card from a deck. For this reason the oldest probability models have only a finite number of sample points. In this chapter we introduce the techniques for computing probabilities of events in a finite process. We will later see that these techniques are useful in many other processes as well.

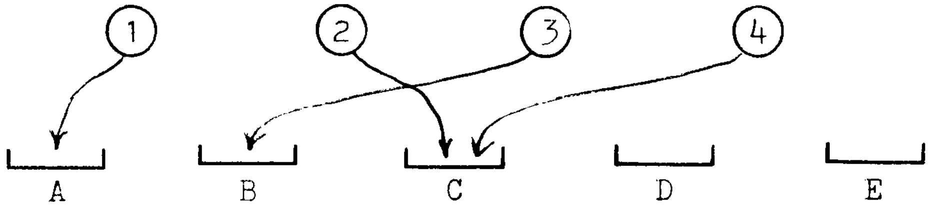
1. The Basic Models

The basic situation for a finite process is the following. Imagine that we have a population consisting of individuals. They may be people, cards or whatever. Now choose one individual from the population. How many ways can we do this? The answer, of course, is the number of individuals in the population. The individual we chose is called a sample (of size 1) from the population. More generally suppose that we choose not just one but a whole sequence of k individuals from the population. The sampling procedure we envision here is successive: we choose one individual at a time. How many ways can we do this? The answer will depend on whether the same individual can be chosen more than once, or equivalently on whether a chosen individual is returned to the population before another is chosen. The two kinds of sampling are called sampling with replacement and sampling without replacement. The sequence of chosen individuals is called a sample of size k.

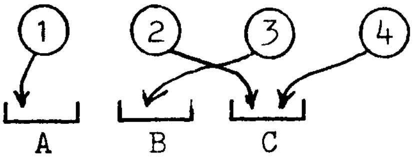
To illustrate this we consider dice and cards. A single roll of a die samples one of its six faces. If we roll the die k times, or equivalently if we roll k dice, we are choosing a sample of size k with replacement from the population of six faces. Similarly, a choice of one card from the standard 52 card deck of cards is a sample from a population of 52 individuals. However, if we deal out k cards, we are choosing a sample of size k without replacement from the population of cards. Note that in a sample the order matters, so that a sample of cards is not the same as a hand of cards, for which the order does not matter.

The description of a finite process given above is called the sampling model. It is by no means the only model or the best model for a given situation. For the rest of this section we consider several other models all of which are mathematically equivalent.

The occupancy model is the following. We have k balls or marbles and n boxes. Denote the set of balls by B and the set of boxes by U (for urns). A placement is a way of placing the balls into the boxes, each ball in some box. For example, here is a placement of 4 balls into 5 boxes:



In the distribution model we have an alphabet U whose members are called letters. A word of length k is any sequence of k letters from the alphabet. The distribution model is easily seen to be equivalent to the occupancy model: the letters correspond to the boxes and the positions of the letters in a word correspond to the balls.



ACBC

A placement of
4 balls into 3 boxes

The corresponding
4 letter word.

If we regard the alphabet U as a population whose individuals are letters, then it is easy to see that a word is just a sample with replacement. Therefore the distribution and occupancy models are both equivalent to the sampling model with replacement. In terms of the occupancy model, sampling without replacement means that no box has more than one ball. In terms of the distribution model, this means that no letter appears twice in a word.

The Mathematical Model. In mathematics a placement of balls is called a function from B to U. Sampling without replacement corresponds to one-to-one functions.

The Physics Model. In physics the balls are called particles, and the boxes are called states, while a placement is called a configuration or system. For a given placement, the occupation number of box i , called Θ_i, is the number of balls placed in box i. The occupation numbers trivially satisfy $\sum_{i=1}^{n} \Theta_i = k = |B|$. For example saying that the placement corresponds to a one-to-one function (or sampling without replacement) is the same as saying Θ_i equals 0 or 1 for all i . In physics such a restriction is called an exclusion principle: a given state may have at most one particle in it. The physics model we have described here goes by the name of Maxwell-Boltzman statistics (with or without the exclusion principle). We shall see other statistics in later sections, which are more physically realistic.

At this point we can make a dictionary for translating terms from one model to another.

Model	Terminology	(terms in one column are equivalent)	
Occupancy	placement	balls	B
Distribution	word	places	-
Sampling	ordered sample	position	-
Mathematics	function	-	domain
Physics*	configuration	particles	-
Astrology	horoscope	planets	solar system

Model	Terminology		
Occupancy	boxes	U	at most one ball per box
Distribution	letters	alphabet	no repeated letters
Sampling	individuals	population	without replacement
Mathematics	-	range	one-to-one function
Physics*	states	-	exclusion principle
Astrology	signs	Zodiac	horoscope

* Maxwell-Boltzman statistics

2. Counting Rules and Stirling's Formula

Placements in their many variations and equivalent forms are the most commonly encountered objects in probability computations. A roll of dice, a hand of cards, even a configuration of particles in chemistry or physics are all forms of placements. In this section we will concentrate on the most basic rules for counting collections of placements. In section four we will consider the more subtle kind of counting necessary in the atomic and sub-atomic domains of chemistry and physics.

The First Rule of Counting

The most fundamental rule of counting is one so obvious that it doesn't seem necessary to dwell on it:

First Rule of Counting. If an object is formed by making a succession of choices such that there are

n_1 possibilities for the first choice

n_2 possibilities for the second choice

etc.

Then the total number of objects that can be made by making a set of choices is

$$n_1 \times n_2 \times n_3 \times \cdots .$$

Note that by a "succession of choices" we mean that after the first choice is made, there are n_2 choices for the second choice, and similarly for subsequent choices.

We illustrate this rule with some familiar examples.

Throwing Dice. How many ways can we roll three dice? A roll of three dice requires three "choices": one for each die . Since each die has six possibilities, there are $6^3 = 216$ rolls of three dice.

Notice that it does not matter whether we view the three dice as being rolled one at a time or all at once. We will call such choices independent: the rolls of the other dice do not affect the set of possibilities available to any given die.

Dealing a Poker Hand. How many ways can we deal five cards from a deck of 52 cards? **(We consider as different two hands** having the same cards but dealt in different orders.) A deal of five cards consists of five choices. There are 52 choices for the first card, 51 choices for the second card, etc. The total number of deals is then $52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 = 311,875,200$. Unlike the situation for rolls of three dice, the cards dealt in the above example are not independent choices. The earlier cards we dealt do affect the set of possibilities for later cards. However the earlier cards do not affect the number of possibilities for a given later deal of a card. Hence the first rule still applies.

Before we consider more complex counting problems we restate the above two examples in the general language of distribution and occupancy.

Arbitrary Placements. The total number of ways to place k balls into n boxes or equivalently the total number of k-letter words made from an alphabet of n letters is n^k: each ball or letter can be chosen in n ways independently of the others.

Placements no two in one box. The total number of ways to place k balls into n boxes so that no two occupy the same box or equivalently the total number of k-letter words made from an alphabet of n letters and having no repeated letters is

$$(n)_k = n(n-1)\dots(n-k+1)$$

There are k factors in this product, one for each choice. This product is called the lower or falling factorial.

An important special case of the second formula is the one for which k = n. Such a placement has a special name: it is a permutation. For example, if we deal all 52 cards from a standard deck, there are $(52)_{52} = 52\cdot 51\cdot 50\cdots 3\cdot 2\cdot 1$ ways for this to happen. This is a very large number, and we will discuss techniques for approximating it below.

Permutations occur so frequently in computations that we have a special notation for the total number of them.

__Definition.__ The total number of ways to place n balls into n boxes, each ball in a different box, or equivalently the number of n-letter words using all the letters from an n-letter alphabet is called __n-factorial__ and is written

$$n! = (n)_n = n(n-1)\ldots 3\cdot 2\cdot 1.$$

Arbitrary placements of k balls into n boxes	n^k
Placements no two in one box of	
k balls into n boxes	$(n)_k$
n balls into n boxes	$n!=(n)_n$

Table 1: Placements

__Stirling's Formula.__

The computation of factorials is so common in probability that it is a great relief to learn that there is an easy way to compute them. The method makes use of an approximation known as Stirling's Formula. The precise mathematical statement is the following:

$$\lim_{n\to\infty}\left(\frac{n^n e^{-n}\sqrt{2\pi n}}{n!}\right) = 1,$$

but in practice this is what one uses:

$$\boxed{n! \simeq n^n e^{-n}\sqrt{2\pi n}}$$

Stirling's Formula

The symbol "≃" means "approximately equal to", and in practice in any expression involving large factorials, we replace each factorial by the right-hand side of Stirling's Formula above.

For example, the total number of permuatations of a standard deck of cards is approximately:

$$52^{52}e^{-52}\sqrt{104\pi} \simeq 8.053\times 10^{67}$$

The Second Rule of Counting.

Poker Hands. Anyone who has played cards knows that one normally does not care about the order in which one is dealt a hand of cards. That is, a hand of cards is not the same as a deal of cards. A poker hand is defined to be a subset of five cards from a standard deck, and the order in which the cards are obtained is immaterial. We cannot count the number of poker hands using the first rule of counting because they violate the fundamental premise of that rule: the object must be obtained by successive choices. However, every poker hand can be dealt in precisely 5! = 120 ways. Therefore the number of poker hands is

$$\frac{52\cdot51\cdot50\cdot49\cdot48}{5!} = \frac{311,875,200}{120} = 2,598,960$$

This illustrates the second rule of counting, also called the "shepherd principle": if it is too difficult to count sheep, count their legs and then divide by four.

Second Rule of Counting. If we wish to count a set of objects obtained by making choices but for which the order of choice does not matter, count as if it did matter and then divide by a factor: the number of ordered objects per unordered object.

Let us illustrate this in a more complicated situation.

Bridge Games. How many bridge situations are possible? By definition a bridge game consists of four people being dealt 13 cards each. However, the order in which each of the four hands is dealt does not matter. So we first count as if the order did matter: this gives 52! possible deals. But each hand can be dealt 13! ways, and there are four hands for a total of $(13!)^4$ ways to deal a given bridge game. Therefore there are

$$\frac{52!}{(13!)^4} \simeq 5.36447 \times 10^{28}$$

possible bridge situations. The symbol "≃" means "approximately equal to."

One must be careful when applying the second rule to make certain that the number of ordered objects is the same for any unordered object. In Section 4 we will give an example of this kind of difficulty. Meanwhile you are now equipped to perform all the (unstarred) counting computations in the exercises.

Before we end this section we cannot resist one more generalization suggested by the above bridge example. Using the language of the occupancy model, a bridge game consists of 52 balls being placed in four boxes such that every box has exactly 13 balls. More generally, the number of ways that k balls can be placed in n boxes so that θ_1 balls are in the first box, θ_2 balls are in the second box, etc. is, by the second rule of counting,

$$\frac{k!}{\theta_1!\theta_2!\ldots\theta_n!}$$

Note that for the above formula to make sense we must have $\theta_1 + \theta_2+\ldots+\theta_n = k$. This expression, called the multinomial coefficient, is written

$$\binom{k}{\theta_1,\theta_2,\ldots\theta_n} = \frac{k!}{\theta_1!\theta_2!\ldots\theta_n!} ,$$

and is prounced "k choose $\theta_1,\theta_2,\ldots,\theta_n$." The numbers $\theta_1,\theta_2,\ldots,\theta_n$ are called the occupation numbers of the placement.

An important special case of the multinomial coefficient is the case n = 2, when it is called the binomial coefficient. This number should be a familiar concept from the binomial expansion in algebra and calculus. Because a placement of balls into two boxes is completely determined by the choice of those in one of the boxes, we can also interpret the binomial coefficient $\binom{k}{\theta_1,\theta_2}$ as the number of θ_1-element subsets of a k-element set. The binomial coefficient is often abbreviated to $\binom{k}{\theta_1} = \binom{k}{\theta_1,\theta_2}$, and is then pronounced simply "k choose θ_1". Using this notation we can quickly compute the number of poker hands because a poker hand consists of five cards chosen from 52 cards:

$$\binom{52}{5} = \binom{52}{5,47} = \frac{52!}{5!47!} = \frac{(52)_5}{5!}$$

Furthermore, we can use binomial coefficients and the first rule of counting to find another formula for the multinomial coefficient. A placement of k balls into n boxes with occupation numbers $\theta_1,\theta_2,\ldots,\theta_n$ can be made using n-1 choices: choose θ_1 balls from the k balls and put them in the first box, choose θ_2 balls from the remaining $k-\theta_1$ balls and put them in the second box,..., choose θ_{n-1} balls from the remaining $k-\theta_1-\ldots-\theta_{n-2}$ balls and put them in the next to last box, and finally put the last $\theta_n = k-\theta_1-\ldots-\theta_{n-1}$ balls in the last box (no more choice is necessary). Therefore:

$$\binom{k}{\theta_1,\theta_2,\ldots,\theta_n} = \binom{k}{\theta_1}\binom{k-\theta_1}{\theta_2}\cdots\binom{k-\theta_1-\ldots-\theta_{n-2}}{\theta_{n-1}}$$

Placements of k balls into n boxes with occupation numbers $\theta_1,\ldots,\theta_n$	$\binom{k}{\theta_1,\theta_2,\ldots,\theta_n}$
Subsets of size θ of a set of k balls	$\binom{k}{\theta} = \binom{k}{\theta,k-\theta}$
Relationship between multinomial and binomial coefficients	
$\binom{k}{\theta_1,\theta_2,\ldots,\theta_n} = \binom{k}{\theta_1}\binom{k-\theta_1}{\theta_2}\cdots\binom{k-\theta_1-\ldots-\theta_{n-2}}{\theta_{n-1}}$.	

Table 2: Multinomial and Binomial Coefficients

3. Computing Probabilities.

Consider a finite sample space Ω. The events of Ω are the subsets $A \subseteq \Omega$. We would like to see what a probability measure P on Ω means. Remember that P is defined on event~~s~~ of Ω. An event $A \subseteq \Omega$ may be written as a finite set

$$A = \{\omega_1,\omega_2,\ldots,\omega_n\}$$

and by additivity any probability measure P must satisfy

$$P(A) = P(\{\omega_1\}) + P(\{\omega_2\})+\ldots+P(\{\omega_n\}).$$

We call the events $\{\omega\}$, having just one outcome $\omega \in \Omega$, the atoms of Ω. It is important to distinguish between an outcome and an atom: the atom $\{\omega\}$ of an outcome ω is the event "ω occurs". The distinction is roughly the same as that between a noun and a simple sentence containing it, e.g., between "tree" and "This is a tree."

What we have just shown above is that *every probability measure* P *on a finite sample space* Ω *is determined by its values on atoms*. The value on an arbitrary event $A \subseteq \Omega$ is then computed by the formula:

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

The values of P on the atoms may be assigned arbitrarily so long as:

(1) For every atom $\{\omega\}$, $0 \leq P(\{\omega\}) \leq 1$,

(2) $\sum_{\omega \in \Omega} P(\{\omega\}) = 1$.

Whenever (1) and (2) hold, P defines a consistent probability measure on Ω.

The simplest example of a probability measure on a finite sample space Ω is the one we will call the *equally likely probability*; it is the unique probability measure P on Ω for which every atom is assigned the same probability as any other. Hence for every atom $\{\omega\}$, $P(\{\omega\}) = 1/|\Omega|$. For more general events $A \subseteq \Omega$, this probability has a simple formula:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{no. of outcomes } \omega \in A}{\text{total no. of outcomes in } \Omega}$$

The equally likely probability is quite common in gambling situations as well as in sampling theory in statistics, although in both cases great pains are taken to see to it that it really is the correct model. For this reason this probability has something of an air of artificiality about it, even though it or probability measures close to it do occur often in nature. Unfortunately nature seems to have a perverse way of hiding the proper definition of Ω from casual inspection.

The phrases "completely at random" or simply "at random" are used to indicate that a given problem is assuming the equally likely probability. The latter phrase is misleading because every probability measure defines a concept of randomness for the sample space in question. Even certainty for one outcome is a special case of a probability measure. In Chapter VII we will justify the use of the description "completely random" for the equally likely probability measure. For now we consider some concrete examples of this probability measure.

Rolling Dice. What is the probability that in a roll of three dice no two show the same face? We already computed $|\Omega|$ in the last section: $|\Omega| = 216$. The event A in question is "the faces of the three dice are all different." We think of an outcome in A as a placement of three balls into 6 boxes so that no two balls are in the same box. There are $(6)_3 = 6 \cdot 5 \cdot 4 = 120$ placements with this property. Hence

$$P(A) = \frac{(6)_3}{6^3} = \frac{120}{216} = .555\ldots$$

Birthday Coincidences. If n students show up at random in a classroom, what is the probability that at least two of them have the same birthday? In order to solve this problem we will make some simplifications. We will assume that there are only 365 days in every year; that is, we ignore leap years. Next we will assume that every day of the year is equally likely to be a birthday. Both of these are innocuous simplifications. Much less innocuous is the assumption that the students are randomly chosen with respect to birthdays. What we mean is that the individual students' birthdays are independent dates.

Let B be the event in question, and let $A = B^c$ be the complementary event "no two students have the same birthday." Now just as we computed in the dice rolling problem above, $P(A) = \dfrac{(365)_n}{365^n}$, and hence $P(B) = 1 - \dfrac{(365)_n}{365^n}$. These probabilities are easily computed on a hand calculator. Here are some values:

$$\begin{aligned} n &= 20, & P(A) &= 0.5886, & P(B) &= 0.4114 \\ n &= 22, & P(A) &= 0.5243, & P(B) &= 0.4757 \\ n &= 25, & P(A) &= 0.4313, & P(B) &= 0.5687 \\ n &= 30, & P(A) &= 0.2937, & P(B) &= 0.7063 \end{aligned}$$

So in a class of 30 students the odds are 7 to 3 in favor of at least two having a common birthday.

Random Committees. In the U.S. Senate a committee of 50 senators is chosen at random. What is the probability that Massachusetts is represented? What is the probability that every state is represented?

In any real committee the members are not chosen at random. What this question is implicitly asking is whether random choice is "fair" with respect to two criteria of fairness. Note that the phrase "at random" is ambiguous. A more precise statement would be that every 50-senator committee is as probable as any other.

We first count $|\Omega|$, the number of possible committees. Since 50 senators are being chosen from 100 senators, there are

$$|\Omega| = \binom{100}{50}$$

committees. Let A be the event "Massachusetts is not represented." The committees in A consist of 50 senators chosen from the 98 non-Massachusetts senators. Therefore,

$$|A| = \binom{98}{50} .$$

Hence $P(A) = \binom{98}{50} \Big/ \binom{100}{50} = \frac{(98)_{50}}{50!} \Big/ \frac{(100)_{50}}{50!} = \frac{(98)_{50}}{(100)_{50}}$

$$= \frac{98 \cdot 97 \cdot \ldots \cdot 49}{100 \cdot 99 \cdot 98 \cdot \ldots \cdot 51} = \frac{50 \cdot 49}{100 \cdot 99} \simeq 0.247.$$

So the answer to our original question is that Massachusetts is represented with probability 1 - 0.247 = 0.753 or 3 to 1 odds. This seems quite fair.

Now consider the event A = "every state is represented." Each committee in A is formed by making 50 independent choices from the 50 state delegations. Hence $|A| = 2^{50}$ and so

$$P(A) = 2^{50} \Big/ \binom{100}{50} \simeq 10^{-14},$$

i.e., essentially impossible. By this criterion random choice is not a fair way to choose a committee.

We computed the above probability by using Stirling's Formula as follows:

$$\frac{2^{50}}{\binom{100}{50}} = \frac{2^{50}}{\left(\frac{100!}{50!50!}\right)} = \frac{2^{50}\cdot(50!)^2}{100!} \simeq \frac{2^{50}(50^{50}e^{-50}\sqrt{2\pi\cdot 50})^2}{100^{100}e^{-100}\sqrt{2\pi\cdot 100}}$$

$$= \frac{2^{50}\cdot 50^{100}\cdot e^{-100}\cdot 2\pi\cdot 50}{100^{100}\cdot e^{-100}\cdot\sqrt{2\pi}\cdot 10} = 2^{50}\cdot\left(\frac{50}{100}\right)^{100}\cdot\sqrt{2\pi}\cdot 5$$

$$= 2^{50}\cdot(\tfrac{1}{2})^{100}\cdot 5\sqrt{2\pi} = 2^{-50}\cdot 5\sqrt{2\pi} \simeq 10^{-15}\cdot 10 = 10^{-14}.$$

The last approximation above is quite rough. We used only that $2^{10} \simeq 1000$ and that π is about 3. All we required was the order of magnitude of the answer. Using a calculator one gets the more exact answer:

$$1.113 \times 10^{-14}$$

Compare this with the following answer obtained (with much more effort) without using Stirling's Formula and correct to 5 decimal places:

$$1.11595 \times 10^{-14}$$

4.* Indistinguishability.

If we roll three dice, the number of possible outcomes is the number of placements of three balls in six boxes: 216. Suppose now that the balls are photons and the boxes are six possible states. Now how many placements are there? The answer is rather surprising: only 56. If we consider electrons instead of photons the answer is even smaller: 20. Moreover, if the six possible states have the same energy, then the 56 states for photons are equally likely. The fact that subatomic particles do not behave as tiny hard balls is one of the major discoveries of physics in this century. The counting problems one encounters in physics are more difficult than those of the previous sections, but a deep understanding of the physics of subatomic particles requires the concepts we present here.

The reason that photons do not behave as dice or balls is a consequence of a propetty known as indistinguishability. In other words, if two photons were interchanged but the rest of the configuration is left unchanged, then the new configuration is identical to the old one. Moreover, given a set of various possible different configurations, all having the same energy, each is as likely to occur as any other. For simplicity, suppose that we have two photons and three states. There are 6 possible configurations:

number of photons in state #1	in state #2	in state #3
2	0	0
1	1	0
1	0	1
0	2	0
0	1	1
0	0	2

Particles which are indistinguishable and for which any number of particles can occupy a given state are called bosons, and we say they obey Bose-Einstein statistics. For example, photons and hydrogen atoms are bosons.

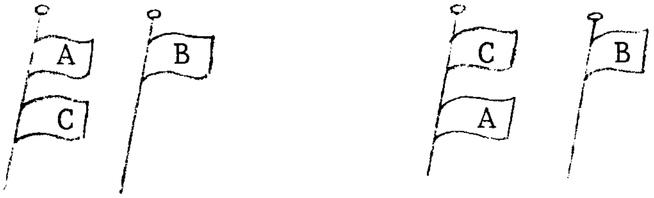
Electrons differ from photons in that two electrons cannot have the same state. As a result, if our configuration consists of two electrons occupying three states, there are only three possible configurations. The fact that two electrons cannot occupy the same state is called the Pauli exclusion principle. Particles which are indistinguishable and which obey the Pauli exclusion principle are called fermions, and we say they obey Fermi-Dirac statistics. For example, electrons, neutrons and protons are fermions.

Fermi-Dirac Statistics: Subsets.

We first count Fermi-Dirac configurations. A Fermi-Dirac configuration means simply that certain of the states are occupied (or "filled") and the rest are not. Thus a Fermi-Dirac configuration is a subset of the states. Since there are n states and k particles, there are $\binom{n}{k}$ possible configurations.

Bose-Einstein statistics: multisets.

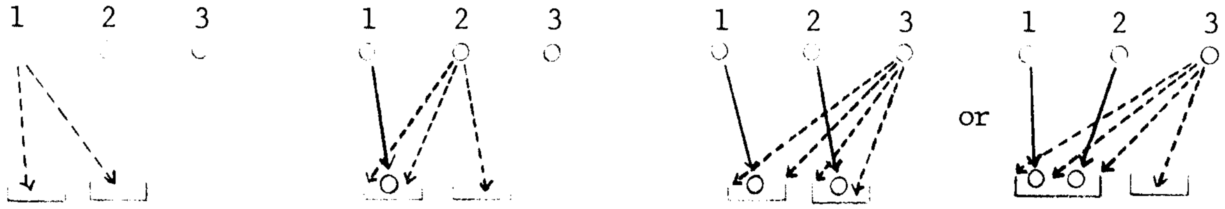
In order to count placements obeying Bose-Einstein statistics, we will use the second rule of counting. However the ordered object corresponding to these placements is a new concept: the disposition. A disposition if k balls in n boxes is a placement together with the additional information of an arrangement in some order of the balls placed in each box. Another model for a disposition is a set of k flags arranged on n flagpoles.



Two different dispositions of three flags on two flagpoles.

Yet another model is that of a disposition of k cars in n traffic lanes on a turnpike.

We can count the number of dispositions using the first rule of counting. The first ball can be placed n ways. The second ball has n+1 choices: either we place it in an unoccupied box (n-1 choices) or we place it before or after the first ball in the occupied box (2 choices). The third ball has n+2 choices. If there are n-2 unoccupied boxes, we can place the third ball in two ways into each of the two occupied boxes. If there are n-1 unoccupied boxes, we can place the third ball in three ways into the occupied box. In general each newly placed ball creates one more "box" for the next ball.



First ball second ball third ball

By the first rule of counting, there are $n(n+1)\ldots(n+k-1)$ dispositions of k balls in n boxes. We call this the <u>rising factorial</u> or $n^{(k)}$. As with the lower factorial, the k denotes the number of factors.

We now consider an alternative way to count dispositions. We first specify the occupancy numbers, <u>then</u> count all dispositions for the given set of occupancy numbers. To see this better, we consider the example of three balls into 2 boxes. There are four choices for the occupancy numbers:

θ_1	θ_2
3	0
2	1
1	2
0	3

Possible occupancy numbers for placing 3 balls into 2 boxes

We now enumerate the dispositions for each set of occupancy numbers:

$\theta_1 = 3$, $\theta_2 = 0$		$\theta_1 = 2$, $\theta_2 = 1$		$\theta_1 = 1$, $\theta_2 = 2$		$\theta_1 = 0$, $\theta_2 = 3$	
123		12	3	1	23		123
213		21	3	2	13		213
132		13	2	1	32		132
312		31	2	3	12		312
231		23	1	2	31		231
321		32	1	3	21		321

The dispositions of 3 balls into 2 boxes.

A glance at the table reveals the following general fact: the number of dispositions of k balls into n boxes having a given set of occupation numbers is precisely $k! = (k)_k$, the number of permutations of k balls.

By the second rule of counting, the number of sets of occupation numbers is $\frac{n^{(k)}}{k!}$. Now a Bose-Einstein configuration means that each state has a certain number of particles in it (possibly none). Thus a Bose-Einstein configuration is nothing but a set of occupation numbers. We will write $\left\langle {n \atop k} \right\rangle$ for $\frac{n^{(k)}}{k!}$ by analogy with the binomial coefficient and call it the <u>multiset</u> <u>coefficient</u>. We use this name because we may interpret a Bose-Einstein configuration as being a multiset: a set of elements together with a nonnegative multiplicity for each element. For example, {a,c,a,a,c,f,g} is a multiset and not a set. The classical terminology for multiset is combination.

Monomials. Consider for example how many monomials of degree 16 can be made with 10 variables $x_1,\ldots,x_{10}$. Each such monomial is a product of 16 x_i's where some must be repeated and the order does not matter. Thus each is a 16-multisubset of the set $\{x_1,\ldots,x_{10}\}$. Therefore there are $\left\langle {10 \atop 16} \right\rangle = \frac{10^{(16)}}{16!} = 2{,}042{,}975$ monomials.

Particles. What is the probability that a random configuration of k particles in n states will have occupation numbers $\theta_1,\ldots,\theta_n$? The answer will depend on the "physics" of the problem, i.e., which statistics is to be used: Maxwell-Boltzmann, Fermi-Dirac or Bose-Einstein. Let A be this event.

(1) Maxwell-Boltzmann. $P(A) = \binom{k}{\theta_1,\ldots,\theta_n} n^{-k}$

(2) Fermi-Dirac. $P(A) = \begin{cases} 0 & \text{if one of the } \theta_i \text{ is greater than one} \\ \binom{n}{k}^{-1} & \text{otherwise} \end{cases}$

(3) Bose-Einstein. $P(A) = \left\langle {n \atop k} \right\rangle^{-1}$

We try two examples: 3 particles in 5 states with occupation numbers (1,0,1,1,0) and (0,3,0,0,0).

	(1,0,1,1,0)	(0,3,0,0,0)
(1) Maxwell-Boltzmann	6/125 = 0.048	1/125 = 0.008
(2) Fermi-Dirac	1/10 = 0.100	0
(3) Bose-Einstein	1/35 ≃ 0.029	1/35 ≃ 0.029

Notice how Bose-Einstein statistics "enhances" the probability of having multiple particles in a single state relative to Maxwell-Boltzmann statistics.

Arbitrary placements	$\left\langle {n \atop k} \right\rangle = \frac{n^{(k)}}{k!}$
Placements with at most one ball per box	$\binom{n}{k} = \frac{(n)_k}{k!}$

Table 3: Placements of k indistinguishable balls into n boxes.

5*. Identities for Binomial and Multiset Coefficients.

The coefficients $\binom{n}{k}$ and $\left\langle {n \atop k} \right\rangle$ satisfy a wealth of identities. Although these can be proved using the formulas, they can often be given combinatorial proofs. That is, we can prove them using only their definitions in terms of "balls into boxes."

1. $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$. To prove this we "mark" one of the boxes, say the last one. Let:

Ω = the set of all k-subsets of the n-set U

B = the set of all k-subsets of the n-set U which contain the last box

C = the set of all k-subsets of the n-set U which do not contain the last box.

Ω consists of all (unordered) k element samples from a population U of size n. B and C are the events "the sample has the last individual" and "the sample does not have the last individual." As events it is clear that $\Omega = B \cup C$ and $B \cap C = \emptyset$. Therefore $|\Omega| = |B| + |C|$. Each of these is easy to count:

$|\Omega| = \binom{n}{k}$ by the definition of the binomial coefficient.

$|B| = \binom{n-1}{k-1}$ since each subset in B consists of the last box together with an arbitrary k-1 element subset of the other n-1 boxes.

$|C| = \binom{n-1}{k}$ since each subset in C consists of a k element subset of the other n-1 boxes.

Remember that Ω, B and C are all sets of sets. So $B \cap C = \emptyset$ means that none of the sets in B are also in C. It does not mean that the sets in B are disjoint from those in C.

Take for example n = 4, k = 3. The set of boxes is U = {1,2,3,4}. The events Ω, B and C are:

$\Omega = \{\{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}\}$

$B = \{\{1,2,4\}, \{1,3,4\}, \{2,3,4\}\}$

$C = \{\{1,2,3\}\}$

Then $|\Omega| = \binom{4}{3} = 4$, $|B| = \binom{3}{2} = 3$, $|C| = \binom{3}{3} = 1$. Be careful to distinguish $\{2,3,4\}$ from $\{\{2,3,4\}\}$, as for example the former has three elements but the latter has only one. Also be careful to distinguish subset <u>of</u> (as in "subset of U") from subset <u>in</u> (as in "subset in B").

2. $\left\langle {n \atop k} \right\rangle = \left\langle {n \atop k-1} \right\rangle + \left\langle {n-1 \atop k} \right\rangle$. This is the multiset analogue of identity 1. We prove it similarly. $\left\langle {n \atop k-1} \right\rangle$ is the number of k-multisubsets of the n boxes which contain <u>at least one</u> copy of the last box: every such multisubset is obtainable by choosing an arbitrary $(k-1)$-multisubset of the n boxes and then throwing in one more copy of the last box. $\left\langle {n-1 \atop k} \right\rangle$ is the number of k-multisubsets of the n boxes which contain <u>no</u> copy of the last box.

Identity 1 gives rise to Pascal's triangle. Because of identity 2, multiset coefficients may also be arrayed in a Pascal-like triangle.

3. $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n-1}{k} + \ldots + \binom{k}{k}$. We prove this by classifying the $(k+1)$-subsets of a set U of $n+1$ boxes according to which is the lowest numbered box in the subset. We assume that U consists of boxes numbered $1,2,\ldots,n+1$. Let

A = the set of all $(k+1)$-subsets of U

B_1 = the set of all $(k+1)$-subsets of U which contain 1

B_2 = the set of all $(k+1)$-subsets of U which contain 2 but not 1

. . .

B_ℓ = the set of all $(k+1)$-subsets of U which contain ℓ but not $1,2,\ldots,\ell-1$

. . .

Then $A = B_1 \cup B_2 \cup \ldots$ and $B_i \cap B_j = \emptyset$ (if $i \neq j$). Therefore $|A| = |B_1| + |B_2| + \ldots$. Each of these is easy to count: $|A| = \binom{n+1}{k+1}$; $|B_1| = \binom{n}{k}$ since every element of B_1 consists of box 1 together with a k-subset of the boxes numbered $2,3,\ldots,n+1$; $|B_2| = \binom{n-1}{k}$; and so on.

4. $\genfrac{\langle}{\rangle}{0pt}{}{n+1}{k} = \genfrac{\langle}{\rangle}{0pt}{}{n}{0} + \genfrac{\langle}{\rangle}{0pt}{}{n}{1} + \ldots + \genfrac{\langle}{\rangle}{0pt}{}{n}{k}$. This is the multiset analogue of identity 3. We classify the k-multisubsets of an $(n+1)$-set according to how many copies of the last element of the $(n+1)$-set are in the multisubset.

5. $\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}$. This one is easiest to prove directly from the formula; however it does have a combinatorial proof. Suppose we make a table of all k-subsets of an n-set U. For example, take $n = 4$, $k = 3$:

①,	②,	③	
①,	②,		④
①,		③,	④
	②,	③,	④

We use $k\binom{n}{k}$ entries when we make this table. We now count the number of entries in another way. Each element of the n-set U appears $\binom{n-1}{k-1}$ times in this table: once for every way of choosing $k-1$ elements from the remaining $n-1$ elements of U. Therefore there are also $n\binom{n-1}{k-1}$ entries in the table. Hence:

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

6. $\left\langle {n \atop k} \right\rangle = \frac{n}{k}\left\langle {n+1 \atop k-1} \right\rangle$ We leave the multiset analogue as an exercise.

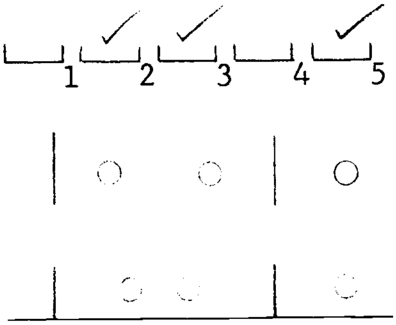
7. $\binom{n}{k} = \binom{n}{n-k}$ Choosing a k-subset is the same as choosing its complement, which has $(n-k)$ elements.

8. $\binom{i+j}{k} = \sum_{\ell=0}^{k} \binom{i}{\ell}\binom{j}{k-\ell}$ Both of these are proved the same way. Start with a set U of $n = i+j$ boxes.

9. $\left\langle {i+j \atop k} \right\rangle = \sum_{\ell=0}^{k} \left\langle {i \atop \ell} \right\rangle \left\langle {j \atop k-\ell} \right\rangle$ Classify every k-element subset according to the number of elements among the first i boxes, in which case $k-\ell$ of the elements will be among the last j boxes. Similarly for multisubsets.

10. $\binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n} = 2^n$ The left hand side counts the number of subsets of an n-set. The right hand side is the number of placements of n distinguishable balls into two boxes, i.e., also counts the number of subsets of an n-set. There is no multiset analogue.

11. $\left\langle {n \atop k} \right\rangle = (-1)^k \binom{-n}{k} = \binom{n+k-1}{k}$ The binomial and multiset coefficients make sense with n any real number. This identity is easily proved explicitly. A combinatorial proof reveals an alternative way to compute the multiset coefficient. The binomial coefficient $\binom{n+k-1}{k}$ counts the number of k-subsets of a set of $n+k-1$ boxes. For a given k-subset of the boxes, charge the boxes in the k-subset to balls and change the remaining $n-1$ boxes into vertical lines marking the boundaries of n new boxes. For example:



the 3-subset consisting of 2, 3, and 5

becomes 3 balls and 2 boundary lines

and thence 3 (indistinguishable) balls in 3 boxes.

Therefore $\binom{n+k-1}{k} = \left\langle {n \atop k} \right\rangle$.

12. $\left\langle {n \atop k} \right\rangle = \left\langle {k+1 \atop n-1} \right\rangle$ Use the above representation of $\left\langle {n \atop k} \right\rangle$, then interchange balls and vertical lines.

6.* Random Integers

It is intuitively obvious that if we choose an integer "at random" it will be even with probability 1/2, divisible by 3 with probability 1/3 and so on. Furthermore, if p and q are different prime numbers, then an integer chosen at random will be divisible by both p and q with probability 1/pq, since it will be so if and only if it is divisible by the product pq. therefore divisibility by p and q are independent events. All this sounds reasonable except that it is not clear how to make sense out of the concept of a "randomly chosen integer" in such a way that every integer is equally likely.

The naive approach is the notion of arithmetic density. Namely let $\Omega = \{1,2,3,...\}$ be the sample space consisting of all integers, and we take the events to be arbitrary subsets of Ω . For an event $A \subseteq \Omega$, the arithmetic density of A is

$$d(A) = \lim_{n\to\infty} \frac{|\{1,2,...,n\} \cap A|}{n}$$

if it exists. For example if D_p is the event "n is divisible by p" or equivalently as a set, $D_p = \{p,2p,3p,...\}$, then it is obvious that $d(D_p) = \frac{1}{p}$. Moreover if p and q are different primes then $d(D_p \cap D_q) = \frac{1}{pq}$. Unfortunately there are several problems with this definition. First of all, d is not defined on all events. Secondly, d is not a probability density even where it is defined. For example if we decompose the set of even integers D_2 into its individual elements

$$D_2 = \{2\} \cup \{4\} \cup \{6\} \cup \ldots$$

we get $d(D_2) = 1/2$ but $d(\{2\}) + d(\{4\}) + \ldots = 0$.

As an example of an event on which d is not defined consider the event "the first digit of n is 1". Call this event F_1. Then

$$F_1 = \{1,10,11,\ldots,19,100,101,\ldots,199,1000,\ldots\}.$$

In this case $\frac{|\{1,\ldots,n\} \cap F_1|}{n}$ forever wanders between $\frac{1}{9}$ and $\frac{5}{9}$ and never "settles down" to any limiting value as $n \to \infty$.

We shall describe a "better" approach to this problem which is nevertheless far from being a complete answer.

Recall from calculus that the series $\sum_{n=1}^{\infty} \frac{1}{n^s}$ converges when $s>1$ and diverges when $s \leq 1$. (The usual way one shows this is the integral test.) The value of this series when $s>1$ is written $\zeta(s)$ and is a famous function called the <u>Riemann zeta function</u>. Computing values of this function is quite difficult, for example $\zeta(2) = \frac{\pi^2}{6}$. However, we shall only need the fact that $\zeta(s)$ exists.

We now define for an event $A \subseteq \Omega$, the <u>Dirichlet density</u> of A with parameter s to be $P_s(A) = \sum_{n \in A} \frac{1}{n^s} \cdot \frac{1}{\zeta(s)}$.

The Dirichlet densities are easily checked to be probability measures. For example $P_s(\Omega) = \sum_{n \in \Omega} \frac{1}{n^s} \cdot \frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{1}{n^s} \frac{1}{\zeta(s)}$

$= \zeta(s) \cdot \frac{1}{\zeta(s)} = 1.$

Let us compute the probability of the event D_p, "n is divisible by p". $P_s(D_p) = \sum_{n \varepsilon D_p} \frac{1}{n^s} \cdot \frac{1}{\zeta(s)}$

$$= (\frac{1}{p^s} + \frac{1}{(2p)^s} + \ldots) \cdot \frac{1}{\zeta(s)} = \frac{1}{p^s} (\frac{1}{1^s} + \frac{1}{2^s} + \ldots) \cdot \frac{1}{\zeta(s)} = \frac{1}{p^s} \cdot \zeta(s) \cdot \frac{1}{\zeta(s)}$$

$= \frac{1}{p^s}$. Similarly, for distinct primes p and q, $P_s(D_p \cap D_q)$ $= \frac{1}{(pq)^s} = P_s(D_p) P_s(D_q)$. Therefore while the Dirichlet density of D_p is not the intuitively expected value $\frac{1}{p}$, it is nevertheless true that being divisible by different primes are independent events.

We find ourselves in a quandary. The notion of arithmetic density is intuitive, but it is not a probability. On the other hand the Dirichlet densities are probabilities, but they depend on a parameter s whose meaning is not easy to explain. Moreover the Dirichlet densities assign the events D_p the "wrong" probability.

We get out of this quandary by a simple expedient: take the limit $\lim_{s \to 1} P_s(A)$. One can prove, although it is very difficult to do so, that if the event A has an arithmetic density, then $d(A) = \lim_{s \to 1} P_s(A)$. Moreover events such as F_1 now have a density, for one can show that $\lim_{s \to 1} P_s(F_1) = \log_{10}(2)$. These probabilities have many useful applications in the theory of numbers.

7. Exercises for

Chapter II Finite Processes

The Basic Models

1. Flip a coin three times. How many ways can this be done? List them. Convert each to the corresponding placement of 3 balls into 2 boxes. Do any of these placements satisfy the exclusion principle?

2. Roll a die twice. List in a column the ways that this can be done. In the next column list the corresponding placements of 2 balls into 6 boxes. Mark the ones which satisfy the exclusion principle. In a third column list the corresponding 2-letter words, using the alphabet {A, B, C, D, E, F}.

3. You are interviewing families in a certain district. In order to ascertain the opinion held by a given family you sample two persons from the family. Recognizing that the order matters in which the two persons from one family are interviewed, how many ways can one sample two persons from a six person family? List the ways and compare with the lists in exercise 2 above. If the two persons are interviewed simultaneously so that order no longer matters, how many ways can one sample two persons from a 6-person family?

4. Return to exercise 2. In a fourth column list the occupation numbers of the six boxes.

The Rules of Counting and Stirling's Formula

5. A small college has a soccer team that plays eight games during its season. In how many ways can the team end its season with five wins, two losses and one tie? Use a multinomial coefficient.

6. Ten students are travelling home from college in Los Angeles to their homes in New York City. Among them they have two cars, each of which will hold six passengers. How many ways can they distribute themselves in the two cars.

The following two problems require a hand calculator.

7. Compute the order of magnitude of 1000!, i.e., compute the integer n for which 1000! is approximately equal to 10^n. [Use a hand calculator and Stirling's Formula to compute the approximate value of $\log_{10}(1000!)$.]

8. How many ways can a 100-member senate be selected from a country having 300 ,000,000 inhabitants?

The Finite Uniform Probability Measure

9. Have the students in your probability class call out their birthdays until someone realizes there is a match. Record how many birthdays were called out. We will return to this problem in Exercise III.25.

10. Give a formula for the probability that in a class of n students at least two have adjacent or identical birthdays.

Ignore leap years. Calculate this probability using a hand calculator for $n = 10, 15, 20$ and 25.

11. Compute the probabilities for two dice to show n points, $2 \leq n \leq 12$. Do the same for three dice.

12. It is said that the Earl of Yarborough used to bet 1000 to 1 against being dealt a hand of 13 cards containing no card higher than 9 in the whist or bridge order. Did he have a good bet? In bridge the cards are ranked in each suit from 2 (the lowest) through 10, followed by the Jack, Queen, King and Ace in this order.

13. May the best team win! Let us suppose that the two teams that meet in the World Series are closely matched: the better team wins a given game with probability 0.55. What is the probability that the better team will win the World Series? Do this as follows. Treat the games as tosses of a biased coin. Express the event "the better team wins" in terms of elementary Bernoulli events, and then compute the probability. We consider in exercise VIII.xx how long a series of games is necessary in order to be reasonably certain that the best team will win.

14. Although Robin Hood is an excellent archer, getting a "bullseye" nine times out of ten, he is facing stiff opposition in the tournament. To win he finds that he must get at least four bullseyes with his next five arrows. However, if he gets five bullseyes, he runs the risk of exposing his identity to the sheriff. Assume that if he wishes to miss the bullseye he can do so with probability 1. What is the probability that Robin wins the tournament?

15. A smuggler is hoping to avoid detection by customs officials by mixing some illegal drug tablets in a bottle containing some innocuous vitamin pills. Only 5% of the tablets are illegal in a jar containing 400 tablets. The customs official tests five of the tablets. What is the probability that he catches the smuggler? [Answer: about 22.7%] Is this a reasonable way to make a living?

16. Every evening a man either visits his mother, who lives downtown, or visits his girl friend, who lives uptown (but not both). In order to be completely fair, he goes to the bus stop every evening at a random time and takes either the uptown or the downtown bus, whichever comes first. As it happens each of the two kinds of buses stops at the bus stop every 15 minutes with perfect regularity (according to a fixed schedule). Yet he visits his mother only around twice each month. Why?

17. In a small college, the members of a certain Board are chosen randomly each month from the entire student body. Two seniors who have never served on the Board complain that they have been deliberately excluded from the Board because of their radical attitudes. Do they have a case? There are 1000 students in the college and the Board consists of 50 students chosen eight times every year.

18. The smuggler of exercise 15 passes through customs with no difficulty even though they test 15 tablets. But upon reaching home he discovers to his dismay that he accidentally put too many illegal drug tablets in with the vitamin pills, for he finds

that 48 of the remaining 385 tablets are illegal. Does he have reason to be suspicious? The question he should ask is the following: given that he packed exactly 48 illegal pills, what is the probability that none of the 15 tested were illegal?

19. Using Stirling's formula, compute the probability that a coin tossed 200 times comes up heads exactly half of the time. Similarly what is the probability that in 600 rolls of a die, each face shows up exactly 100 times?

20. The following is the full description of the game of CRAPS. On the first roll of a pair of dice, 7 and 11 win, while 2, 3 and 12 lose. If none of these occur, the number of dots showing is called the "point," and the game continues. On every subsequent roll, the point wins, 7 loses and all other rolls cause the game to continue. You are the shooter; what is your probability of winning?

21. Compute the probability of each of the following kinds of poker hand, assuming that every five-card poker hand is equally likely. Note that the kinds of hands listed below are pairwise disjoint. For example, in normal terminology a straight hand does not include the straight flush as a special case.

	kind of hand	definition
(a)	"nothing"	none of (b)-(j)
(b)	one pair	two cards of the same rank
(c)	two pair	two cards of one rank and two of another
(d)	three-of-a-kind	three cards of the same rank
(e)	straight	ranks in ascending order (ace may be low card or high card but not both at once)

(f)	full house	three of one rank and two of another
(g)	flush	all cards of the same suit
(h)	straight flush	both (e) and (g)
(i)	four-of-a-kind	four cards of the same rank
(j)	royal flush	(h) with ace high

22. It is an old Chinese custom to play a dice game in which six dice are rolled and prizes are awarded according to the pattern of the faces shown, ranging from "all faces the same" (highest prize) to "all faces different." List the possible patterns obtainable and compute the probabilities. Do you notice any surprises?

23. Some environmentalists want to estimate the number of whitefish in a small lake. They do this as follows. First 50 whitefish are caught, tagged and returned to the lake. Some time later another 50 are caught and they find 3 tagged ones. For each n compute the probability that this could happen if there are n whitefish in the lake. For which n is this probability the highest? Is this a reasonable estimate for n ?

24. A group of astrologers has, in the past few years, cast some 20,000 horoscopes. Consider only the positions (houses) of the sun, the moon, Mercury, Venus, Earth, Mars, Jupiter and Saturn. There are twelve houses in the Zodiac. Assuming complete randomness, what is the probability that at least two of the horoscopes were the same?

25. In a chess championship, a certain number N of games are specified in advance. The current champion must win N games in order to retain the championship, while the challenger must win more than N in order to unseat the champion. The challenger is somewhat weaker than the champion, being able to win only a dozen games out of every 25 games which do not end in a tie. If the challenger is allowed to choose the number N , what should the challenger choose? [Answer: 12]. Reference: Fox, Math. Teacher 54 (1961), 411-412.

26. The three-person duel is a difficult situation to analyze in full generality. We consider just a simple special case. Three individuals, X, Y and Z, hate each other so much they decide to have a duel only one of which can survive. They stand at the corners of an equilateral triangle. The probability of a hit by each of the three participants is 0.5, 0.75 and 1, respectively. For this reason they decide that they will each shoot at whomever they wish, taking turns cyclically starting with X and continuing with Y, then Z, then X again and so on. All hits are assumed to be fatal. What strategy should each employ, and what are their probabilities of survival?

History of Probability

Historically, the modern theory of probability can be said to have begun as a result of a famous correspondence between the mathematicians Blaise Pascal (1623-62) and Pierre de Fermat (1601-65). Their correspondence came about as a result of problems put to Pascal by the Chevalier de Méré, a noted gambler of the time, We give below two of these problems.

27. Two gamblers are playing a game which is interrupted. How should the stake be divided? The winner of the game was to be the one who first won 4 deals out of 7. One gambler has so far won 1 deal and the other 2 deals. They agree to divide the stake according to the probability each had of winning the game, and this probability is to be computed by assuming that each player has equal probability of winning a given game. Express the event that the first gambler wins in terms of elementary Bernoulli events. Then compute the probability. [Answer: 5/16].

28. Chevalier de Méré apparently believed that it is just as probable to show at least one six in four throws of a single die as it is to show at least one double-six in twenty-four throws of a pair of dice. However, de Méré computed the probabilities of these two events and found that one was slightly above, the other slightly below, 0.5. What are the exact probabilities? In exercise IV. 22, we will consider the likelihood that de Méré could have found the distinction between these two probabilities empirically.

29. In what was apparently Isaac Newton's only excursion into probability, he answered a question put to him by Samuel Pepys. The problem was to determine which is more likely, showing at least one six in 6 throws of a die, at least two sixes in 12 throws or at least three sixes in 18 throws. Compute these probabilities and consider the general question of the probability of showing at least n sixes in $6n$ throws of a die.

Indistinguishability

30*. Suppose we have a physical system having three energy levels and two states per energy level (for a total of six states). If two electrons are in the configuration, what is the probability that they occupy the lowest two energy levels (one in each level)? Consider the same question for two photons and two Maxwell-Boltzmann particles. Since the states do not have the same energy, the states are not "equiprobable." Assume that the probability of one particle being in a state is proportional to e^{-E} where E is the energy of the state. In this problem suppose that the three energy levels have respective energies 1, 2, and 3.

In the following problem we admittedly oversimplify a bit too much, but it does illustrate some of the ideas and techniques of modern Physics.

31*. Consider a small piece of metal at ordinary temperatures. It forms a crystal with the nuclei of its atoms appearing in a regular fashion throughout the solid. Most of the electrons may be regarded as being bound to some one nucleus. Some of the electrons, the ones in the outermost orbitals of an atom, have more freedom of movement. Call these the valence electrons. The outermost orbitals of a given atom form an almost continuous band. Let us suppose that the valence electrons act as bosons in this environment with any number being allowed in the set of outermost orbitals of one given atom. Let us suppose also that the outermost orbitals of one given atom all have the same energy.

(Neither of these assumptions is actually true.) Let k be the number of atoms and let n be the number of valence electrons. Compute the distribution of θ_1, the number of valence electrons occupying the outermost orbitals of one specific atom. Now in an actual macroscopic piece of metal, n and k are very large (being on the order of 10^{23}) and so cannot be measured exactly. However, the ratio $\lambda = \frac{n}{k}$ is usually not too difficult to find. Since n and k are so large, we may regard them as being infinite. The distribution of θ_1 is then approximated by letting n and k tend to infinity but in such a way that the ratio $\lambda = \frac{n}{k}$ is held fixed. Find this limiting distribution. Such a distribution can be measured experimentally and used to compute λ as well as to test our model.

Identities

32. Prove formally that $\left\langle {n \atop k} \right\rangle = \frac{n}{k} \left\langle {n+1 \atop k-1} \right\rangle$

33*. Give combinatorial proofs of the above identity as well as the identity $\left\langle {n \atop k} \right\rangle = \left\langle {k+1 \atop n-1} \right\rangle$.

34*. Give a combinatorial proof of the following identity"

$$n\,2^{n-1} = \binom{n}{1} + 2\binom{n}{2} + \cdots + (n-1)\binom{n}{n-1} + n\binom{n}{n} .$$

Random Integers

35* **Compute $\lim_{s\to 1} P_s(F_i)$ for $i = 2,3,\ldots,9$. Now pick out 100 addresses at random from a phone book and tabulate the number having each of the 9 possible first digits (ignore addresses other than natural number addresses). Do these fit with the predicted probabilities? Try 1000 addresses.**

Chapter III Random Variables

A random variable is a new way of answering questions about nature. For example, suppose we toss a coin. How long will it take to get the first head? How can one answer such a question? Sometimes the first head will appear on the first toss, sometimes on the next and so on. Clearly we cannot answer such a question with a single number. The originality of the probabilistic point of view is that it answers such a question with a series of possible answers, each with its own probability.

The intuitive idea of a random variable is that it is the strengthening of the notion of a variable. Recall from calculus and algebra that a variable is a symbol together with a set over which the symbol ranges. For example in calculus one often says "let x be a variable ranging over the real numbers" or more succinctly "let x be a real variable." Now a random variable (or R.V. for short) is a variable together with the probability that it takes each of its possible values.

In particular an integer random variable is a variable n ranging over the integers together with the probability p_n that it takes the value n . Implicit in this is that $\sum_n p_n = 1$, which means that the random variable always takes some value or other. Some of the p_n can be zero, which means that these integers do not occur as values of the random variable. For example, if $p_n = 0$ whenever $n \leq 0$, then the random variable is said to be positive, i.e. it takes only positive integral values.

1. Integer Random Variables

We're now ready for the precise mathematical definition. Don't be surprised if at first this notion doesn't appear to match what we've just been led to expect. It has taken an enormous amount of time and effort to make this notion rigorous so it will require some effort and many examples to make this concept clear.

An integer random variable is a function X defined on a sample space Ω, that takes only integer values. Namely, for every sample point $\omega \varepsilon \Omega$, $X(\omega)$ is an integer. The (probability) distribution of X is the sequence of numbers p_n such that p_n is the probability of the event "X equals n". The event "X equals n" is usually written $(X=n)$. As a subset of Ω , this event is $(X=n) = \{\omega \varepsilon \Omega : X(\omega)=n\}$. We shall generally avoid writing out this set explicitly each time. One should develop an intuitive feeling for the event $(X=n)$.

Of course we have implicitly assumed that the subsets $(X=n)$ really are events of the sample space Ω. This is a technical point that will never be of direct concern to us. Suffice it to say that a fully rigorous definition of an integer random variable is: a function X on a sample space Ω such that its values are all integers and such that the subsets $(X=n)$ are all events of Ω.

The probability distribution of an integer R.V. X always satisfies $p_n \geq 0$ for all n and $\sum_n p_n = 1$. The former property expresses the fact that the p_n are probabilities, while the latter says that X always takes some value.

The intuitive idea of a random variable relates to the precise definition of a random variable in the following way. Whenever we have some measurement with probabilities, look for a sample space and a function on it. The random variable then <u>really</u> comes from observing some phenomenon on this sample space. The fact that we only had a probability distribution at first arose from the fact that we had forgotten about the phenomenon from which the measurement came.

Of course all this means little until we have seen examples.

A.The Bernoulli Process:tossing a coin

Recall that Ω is the set of all infinite sequences of zeros and ones corresponding to the tosses of a biased coin with probability p of coming up heads (or one) and q coming up tails (or zero). Let W_1 be the waiting time for the first head. In other words we ask the question: how long do we have to wait to get the first head? The answer is a probability distribution p_n, where p_n is the probability that we must wait until the n^{th} toss to get the first head. In terms of the terminology of R.V.'s:

$$p_n = P(W_1 = n) \ .$$

How can we compute this? Well, the event $(W_1=n)$ is the event: "at the n^{th} toss we get a head and the preceding n-1 tosses are all tails". In terms of elementary events:

$$(W_1 = n) = T_1 \cap T_2 \cap \dots \cap T_{n-1} \cap H_n \ .$$

Therefore $p_n = P(W_1=n) = q^{n-1}p$.

Just for once let us satisfy ourselves that $\sum\limits_n p_n = 1$:

$\sum\limits_n p_n = \sum\limits_{n=1}^{\infty} q^{n-1}p = p\sum\limits_{n=1}^{\infty} q^{n-1} = p\cdot\frac{1}{1-q} = p\cdot\frac{1}{p} = 1.$ So it checks.

Of course it isn't really necessary that we do this. The very definition of a probability distribution requires that it sum to 1. As we shall see probability theory furnishes a new way to perform some very complicated infinite sums simply by observing that the terms are related to the probability distribution of some integer random variable.

Notice that in understanding W_1 as a random variable we worked completely probabilistically. We never spoke of W_1 as a function. What is W_1 as a function? For each $\omega\varepsilon\Omega$, $W_1(\omega)$ is the first position of the sequence ω such that at that position ω has a 1. For example, $W_1(00011011...)$ is 4. However looking at W_1 as a function is quite unnatural. One should try to think of W_1 purely probabilistically. Indeed, one might say that probability theory gives one a whole new way of looking at sets and functions.

Consider another example. Let W_k be the waiting time for the k^{th} head. The event $(W_k=n)$ is the event: "a head occurs at the n^{th} toss and exactly k-1 heads occur during the preceding n-1 tosses." The probability distribution is:

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

How does one see this? Well, the k-1 heads can occur in any (k-1)-subset of the first n-1 tosses. There are $\binom{n-1}{k-1}$ such subsets. For each such subset, the probability of getting heads in those positions and tails in the others is $p^{k-1}q^{n-k}$. Finally the probability of getting a head on the n^{th} toss is p.

Needless to say it is not very easy to write an explicit expression for the events $(W_k=n)$ in terms of elementary events although that is implicit in our computation above.

Notice too that $\sum_{n=k}^{\infty} \binom{n-1}{k-1} p^k q^{n-k} = 1$, a fact that is not very easy to prove directly.

Consider the event $X_n = \begin{cases} 1 & \text{if } n^{th} \text{ trial is } 1 \\ 0 & \text{if } n^{th} \text{ trial is } 0 \end{cases}$

or more succinctly X_n is the n^{th} trial. The distribution of X_n is

$$p_0 = P(X_n = 0) = q$$

$$p_1 = P(X_n = 1) = p$$

and all other p_n are zero.

Next let S_n be the number of heads in the first n tosses. The distribution of S_n is:

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

because the event $(S_n=k)$ means that k heads and n-k tails occur in the first n tosses. There are $\binom{n}{k}$ ways that the k heads can appear and each pattern has probability $p^k q^{n-k}$ of occurring. The fact that $\sum_k p_k = 1$ is just the binomial theorem:

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1^n = 1 \ .$$

Indeed this is a probabilistic proof of the binomial theorem.

Incidentally the event $(S_n=k)$ is not the same as the event $(W_k=n)$. The distinction is that $(W_k=n)$ requires that there be k heads in the first n tosses and that the k^{th} head occur at the n^{th} toss. $(S_n=k)$ is only the event that k heads occur in the first n tosses. The distinction is reflected in the formulas we found for the distributions in each case.

Another way to represent S_n is:

$$S_n = X_1 + X_2 + \ldots + X_n \ .$$

This illustrates the fact that we may combine random variables using algebraic operations. After all, random variables are functions on Ω and as such may be added, subtracted, etc. Thus if X and Y are integer R.V.'s on Ω , then, as a function,the random variable X+Y takes the value $X(\omega)+Y(\omega)$ on the sample point $\omega\epsilon\Omega$. For example $(W_k=n)$ is the event $(X_1+\ldots+X_{n-1} = k-1)\cap(X_n=1) = (S_{n-1} = k-1)\cap(X_n=1)$. Unfortunately the use of large quantities of symbolism tends to obscure the underlying simplicity of the question we asked. We shall try to avoid doing this if possible.

Now consider the random variable T_k, the length of the gap between the $(k-1)^{st}$ and k^{th} heads in the sequence of tosses.

$$\omega = \underbrace{\underline{0001}}_{T_1=W_1}\ \underbrace{\underline{00001}}_{T_2}\ \underbrace{\underline{1}}_{T_3}\ \underbrace{\underline{001}}_{T_4}\ldots$$

The T_k's and W_k's are related to each other:

$$T_k = W_k - W_{k-1}$$
$$W_k = T_1 + T_2+\ldots+T_k$$

What is the distribution of T_k? When we later have the notion of conditional probability we will have a very natural way to compute this. However we can nevertheless easily compute the distribution of T_k because of the independence of the various tosses of the coin. In other words when computing $P(T_k=n)$ we may imagine that we start just after the $(k-1)^{st}$ head has been obtained. Therefore the distribution of T_k is

$$p_n = P(T_k=n) = P(W_1=n) = q^{n-1}p,$$

exactly the same distribution as that of W_1.

Notice that T_k for $k>1$ is not the same random variable as W_1, and yet their distributions are the same. How can this be? Actually we have already seen this phenomenon before but didn't notice it because it was too trivial an example: $X_1, X_2, \ldots$ are all different random variables, but they all have the same distributions. This phenomenon will occur frequently and is very important.

Definition. Two integer random variables X and Y are said to be equidistributed or stochastically identical when

$$P(X = n) = P(Y = n) \quad \text{for all integers } n \ .$$

Thus for example W_1 and T_k are equidistributed R.V.'s.

Similarly the X_n are equidistributed R.V.'s. Although X_1 and X_2 measure completely different phenomena, they have exactly the same probabilistic structure.

B. The Bernoulli Process: random walk

Consider the random variables X_n' given by:

$$X_n' = \begin{cases} 1 & \text{if the } n^{th} \text{ trial is } 1 \\ -1 & \text{if the } n^{th} \text{ trial is } 0 \end{cases}$$

X_n and X_n' are closely related: $X_n' = 2X_n - 1$. However if we form the random variable analogous to S_n we measure a quite different phenomenon. Let $S_n' = X_1' + \ldots + X_n'$, then S_n' is the position of a random walk after n steps: a step to the right gives +1, a step to the left gives -1, so the sum of the first n steps is the position at that time.

What is the probability distribution of S_n'? This calculation is a good example of a "perturbation" (or change of variables) applied to a model. We want to compute $P(S_n'=x)$. Here we use x for an integer; think of it as a point on the x-axis. Let h be the number of heads and t the number of tails, both during the first n tosses. Then:

$$x = h-t \quad \text{and} \quad n = h+t \ .$$

Solving for h and t gives:

$$h = \frac{1}{2}(x+n) \text{ and } t = \frac{1}{2}(n-x).$$

Therefore:

$$P(S_n' = x) = P(S_n = \frac{1}{2}(x + n)) = \binom{n}{\frac{1}{2}(x+n)} p^{\frac{1}{2}(x+n)} q^{\frac{1}{2}(n-x)} .$$

C. Independence and Joint Distributions

Recall that two events A and B are independent when $P(A \cap B) = P(A)P(B)$. This definition is abstracted from experience; as for example when tossing a coin, the second time the coin is tossed, it doesn't remember what happened the first time. We extend this notion to random variables. Intuitively two random variables X and Y are independent if the measurement of one doesn't influence the measurement of the other. In other words the events expressible in terms of X are independent of those expressible in terms of Y. We now make this precise.

Definition. Two integer random variables X and Y are *independent* when

$$P((X=n_1) \cap (Y=n_2)) = P(X=n_1)P(Y=n_2)$$

for every pair of integers n_1, n_2.

We illustrate this with our standard example: the Bernoulli process. X_k and X_n are independent when $k \neq n$. This is obvious from the definition of the Bernoulli process.

Less obvious is that T_k and T_n are independent when $k \neq n$. We check this for T_1 and T_2. By previous computations,

$$P(T_1=n_1) = q^{n_1-1}p \text{ and } P(T_2=n_2) = q^{n_2-1}p .$$

Now compute $P((T_1=n_1) \wedge (T_2=n_2))$. The event $(T_1=n_1) \wedge (T_2=n_2)$ means that the first n_1+n_2 tosses have precisely the pattern:

$$\underbrace{00...01}_{n_1}\ \underbrace{00...01}_{n_2}$$

Therefore $P((T_1=n_1) \wedge (T_2=n_2)) = q^{n_1-1}p\,q^{n_2-1}p$. Since $P((T_1=n_1) \wedge (T_2=n_2))$ is the same as $P(T_1=n_1)P(T_2=n_2) =$ $q^{n_1-1}pq^{n_2-1}p$, we conclude that T_1 and T_2 are independent.

On the other hand, W_k and W_n are not independent R.V.'s. This is intuitively obvious, but we will check it nevertheless in the case of W_1 and W_2. We previously computed:

$$P(W_1 = n_1) = q^{n_1-1}p$$

$$P(W_2 = n_2) = (n_2-1)q^{n_2-2}p^2 \quad .$$

Now $(W_1=n_1)\wedge(W_2=n_2)$ is the same as the event $(T_1=n_1)\wedge(T_2=n_2-n_1)$, both being the event that the first n_2 tosses have the pattern:

$$\overbrace{\underbrace{00\ldots01}_{n_1}\ \underbrace{00\ldots01}_{n_2-n_1}}^{n_2}$$

Therefore, $P((W_1=n_1)\wedge(W_2=n_2)) = \begin{cases} 0 & \text{if } n_2 \leq n, \\ q^{n_2-2}p^2 & \text{if } n_2 > n_1 \end{cases}$

Since $P((W_1=n_1)\wedge(W_2=n_2)) \neq P(W_1=n_1)\ P(W_2=n_2)$ (in particular when $n_1 \geq n_2 \geq 2$ one side is zero and the other is not), W_1 and W_2 are not independent. In other words W_1 <u>influences</u> W_2.

When two R.V.'s are not independent, is there a way to measure the dependence of one of them on the other? In more common parlance, how do we measure the "correlation" of two R.V.'s? We measure this with the joint distribution of two random variables.

Definition. For two integer random variables X and Y, the joint distribution of X and Y is

$$c_{n_1,n_2} = P((X = n_1) \cap (Y = n_2)) \ .$$

The numbers c_{n_1,n_2} cannot be computed in general from the individual distributions of X and Y. The joint distribution measures the total dependence of X and Y or equivalently the cause and effect of one R.V. on the other.

Joint distributions have the following properties:

(1) $\sum_{n_1} \sum_{n_2} c_{n_1,n_2} = 1$, i.e. something must happen.

(2) $\sum_{n_2} c_{n_1,n_2} = P(X = n_1)$.

(3) $\sum_{n_1} c_{n_1,n_2} = P(Y = n_2)$.

The distributions of X and Y considered relative to their joint distribution are called the marginal distributions or simply the marginals. Despite the fancy terminology, the marginals are simply the distributions of X and Y with which we are already familiar.

Just as we have the joint distribution of two random variables, we can have the joint distribution of any finite collection of random variables. The formulas are so obvious that we won't bother to write them down explicitly.

We now compute some examples. If X and Y are independent random variables with distributions $p_{n_1} = P(X=n_1)$ and $r_{n_2} = P(Y = n_2)$, then their joint distribution is

$c_{n_1,n_2} = P((X = n_1)\wedge(Y = n_2)) = P(X = n_1)\ P(Y = n_2) = p_{n_1}r_{n_2}$.

Therefore the joint distribution of independent R.V.'s is the product of the marginals.

Next consider the random variables W_j and W_k $(j<k)$. Their joint distribution is

$$c_{n_1,n_2} = P((W_j = n_1)\wedge(W_k = n_2)).$$

Of course we must have $n_1<n_2$. The event $(W_j=n_1)\wedge(W_k=n_2)$ means that we have j-1 heads in the first n_1-1 tosses and k-j-1 heads in the "gap" of length n_2-n_1-1 between the j^{th} and k^{th} heads.

$$\overbrace{\underbrace{010\ldots1\underset{\substack{\uparrow\\ j^{th}\\ \text{head}}}{1}}_{n_1}\ 001\ldots.\underset{\substack{\uparrow\\ k^{th}\\ \text{head}}}{1}}^{n_2}$$

Writing all this out gives:

$$c_{n_1,n_2} = P((W_j=n_1)\wedge(W_k=n_2)) = \binom{n_1-1}{j-1}p^j q^{n_1-j}\binom{n_2-n_1-1}{k-j-1}p^{k-j}q^{n_2-n_1-(k-j)}$$

$$=\binom{n_1-1}{j-1}\binom{n_2-n_1-1}{k-j-1}p^k q^{n_2-k}$$

The total number of tosses involved is n_2: exactly k of them are heads and n_2-k are tails. This furnishes a quick check that the exponents on p and on q are correct.

As a final example, we compute the joint distribution of the first k waiting times. For $n_1<n_2<\ldots<n_k$, the joint distribution is:

$$c_{n_1,n_2,\ldots,n_k} = P((W_1=n_1)\wedge(W_2=n_2)\wedge\ldots\wedge(W_k=n_k)).$$

This is actually quite easy to compute because there is only one "way" to get the event $(W_1=n_1)\wedge\ldots\wedge(W_k=n_k)$ up to the n_k^{th} toss. Therefore:

$$c_{n_1,n_2\ldots,n_k} = P((W_1=n_1)\wedge\ldots\wedge(W_k=n_k)) = pq^{n_1-1}pq^{n_2-n_1-1}\ldots=p^kq^{n_k-k}.$$

D*. Fluctuations of Random Walks

Recall that the basic random variables of the Random walk sample space are X_n' for n=1,2,... . These are independent random variables taking values $\pm$ 1 with probability p and q respectively. They represent the direction taken during the n^{th} step of the random walk. The position of the random walk after the n^{th} step is then the random variable $S_n' = X_1' + X_2' + \ldots + X_n'$. We computed the distribution of S_n' in general in section 1. For the special case of a symmetric random walk,

$$P(S_n' = x) = \binom{n}{\frac{n+x}{2}} \frac{1}{2^n} .$$

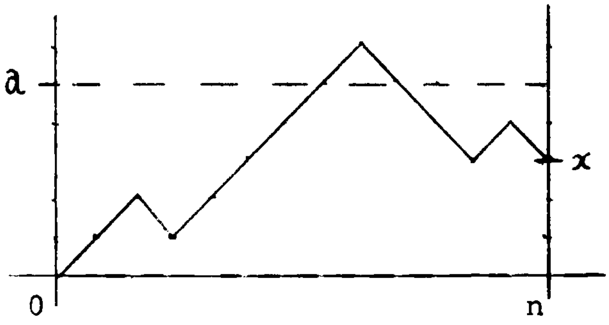
We will write p(n,x) for the above probability.
Note that S_n' takes only even values for even n and only odd values for odd n.

For the rest of this section we will consider only the case of a symmetric random walk, i.e. one for which $p = q = \frac{1}{2}$.

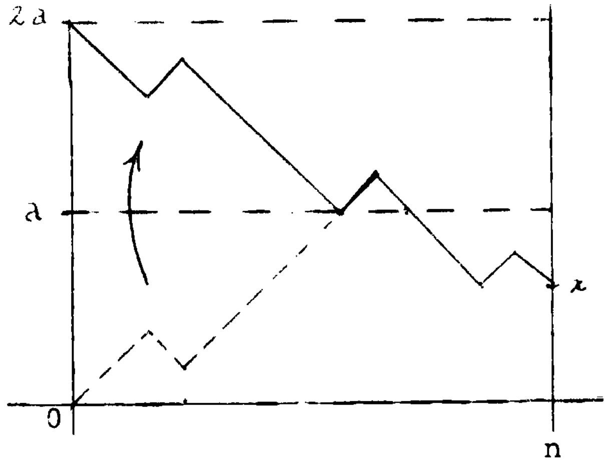
First Passage Time and the Reflection Principle

The event $(S_n' = 0)$ means that the random walk has returned to the origin after n steps. However, it could have returned many times before. When was the first time it returned to the origin (or more generally any point a>0)? We answer this by computing the probability distribution of the random variable T_a, the time when the random walk first encounters the point a , i.e. the first time n such that $S'_n = a$.

To compute the distribution of T_a we use an important principle called the reflection principle. Consider the event $C_{n,a,x}$ = "the random walk is at position x at time n and at some previous time was at position a ". The following is the graph of a typical random walk in $C_{n,a,x}$:



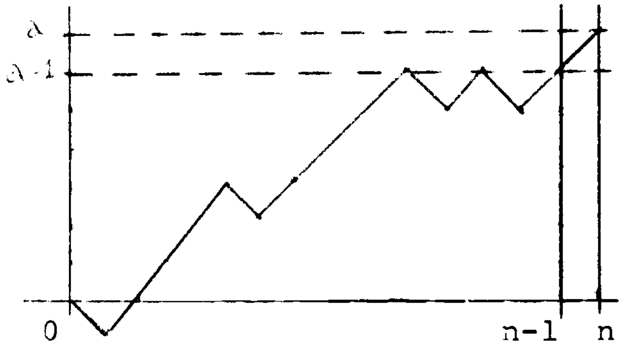
Now observe that every random walk in $C_{n,a,x}$ is necessarily at position a for a first time. We take each random walk in $C_{n,a,x}$ and "pivot" or "reflect" it up to the first time that it reaches position a:



In this way we get a random walk from 2a to x. Conversely, <u>any</u> random walk from 2a to x necessarily crosses a at some time, so <u>every</u> random walk from 2a to x is <u>uniquely</u> determined in this way! Now shift the axis so that 2a becomes the origin and x becomes the point x-2a. Then we conclude that $P(C_{n,a,x})$ is the same as $P(S_n' = x-2a)$. By symmetry this is the same as $P(S_n' = 2a-x)$. Thus

$$P(C_{n,a,x}) = p(n,2a-x).$$

We are now ready to compute $P(T_a=n)$. We first note that $(T_a=n)$ necessarily implies that the random walk moved from a-1 to a at step n. Prior to step n the random walk <u>never</u> achieved position a but ends at position a-1 at step n-1. This is just the "complement" of the event $C_{n-1,a,a-1}$.



[This random walk is at a-1 at time n-1 so it is in $(S'_{n-1}=a-1)$. It never reaches position a so it is not in $C_{n-1,a,a-1}$.]

More precisely it is the difference of events:

$$(S'_{n-1} = a-1) - C_{n-1,a,a-1} .$$

Putting this all together:

$$(T_a = n) = ((S'_{n-1} = a-1) - C_{n-1,a,a-1}) \cap (X'_n = 1).$$

This is the intersection of independent events. Therefore

$$P(T_a=n) = P((S'_{n-1}=a-1) - C_{n-1,a,a-1})\ P(X'_n=1)$$

$$= \frac{1}{2}\, P((S'_{n-1}=a-1) - C_{n-1,a,a-1}) .$$

Now $C_{n-1,a,a-1}$ is a subevent of $(S'_{n-1}=a-1)$ so

$$P(T_a=n) = \frac{1}{2}\,[P(S'_{n-1}=a-1) - P(C_{n-1,a,a-1})]$$

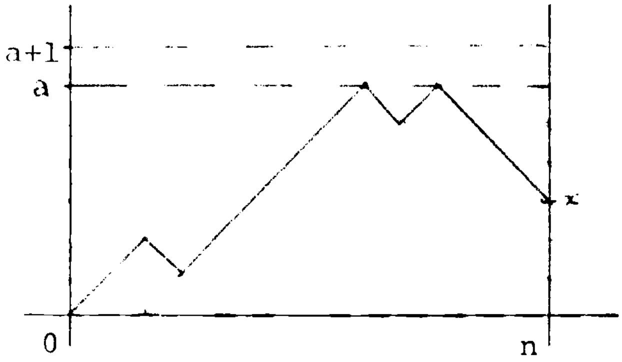
$$= \frac{1}{2}\,[p(n-1,a-1) \;-\; p(n-1,2a-(a-1))]$$

$$= \frac{1}{2}\,[p(n-1,a-1) \;-\; p(n-1,a+1)].$$

Maximum Position

We next ask how far the random walk travels to the right (i.e. the maximum position achieved). Let M_n be this maximum for an n-step random walk. Conveniently, the events $C_{n,a,x}$ are just what we need to compute the distribution of M_n. Namely, we use the same "trick" of subtracting one of the $C_{n,a,x}$; but this time from another event of this kind.

First note that the event $C_{n,a+1,x}$ is a subevent of $C_{n,a,x}$; for if a random walk achieves position a+1, then it must have some time previously been at position a. Thus $P(C_{n,a,x}-C_{n,a+1,x}) = P(C_{n,a,x}) - P(C_{n,a+1,x}) = p(n,2a-x) - p(n,2a+2-x)$. But the event $C_{n,a,x} - C_{n,a+1,x}$ means that the random walk achieved position a but never achieved position a+1, i.e. the maximum achieved by the random



Typical random walk in $C_{n,a,x}-C_{n,a+1,x}$

walk is precisely a. The only distinction between this event and the event $(M_n=a)$ is that the latter does not specify the ending point x specified by the former. So to get $P(M_n=a)$ we add up the $P(C_{n,a,x}-C_{n,a+1,x})$ for all possible values of x. The permissible values of x range from a down to any reachable negative point on the x-axis.

Thus $P(M_n=a)$ is this sum:

$$p(n,2a-a) - p(n,2a+2-a) \qquad (x=a)$$

$$+\ p(n,2a-(a-1)) - p(n,2a+2-(a-1)) \qquad (x=a-1)$$

$$+\ \ldots \qquad (x<a-1)$$

which equals

$$\begin{aligned}
&\ p(n,a) - p(n,a+2)\\
&+ p(n,a+1) - p(n,a+3)\\
&+ p(n,a+2) - p(n,a+4)\\
&+ p(n,a+3) - p(n,a+5)\\
&+ \ldots
\end{aligned}$$

Cancelling in the obvious way, we get:

$$P(M_n=a) = p(n,a) + p(n,a+1).$$

Note that for each a only one of the summands on the right is nonzero.

We summarize the computations in this table.

T_a = time of first passage to or through position a.

$$P(T_a=n) = \frac{1}{2}\,(p(n-1,a-1) - p(n-1,a+1)).$$

M_n = maximum position achieved up to time n.

$$P(M_n=a) = p(n,a) + p(n,a+1).$$

E. Expectations

Definition. Suppose X is an integer R.V. with distribution $p_n = P(X=n)$.

The expectation or mean or average value of X is

$$E(X) = \sum_n n \cdot p_n = \sum_n n \cdot P(X = n) \ .$$

It can happen that this sum does not exist. We won't worry about this. Implicit in any statement about expectations is the assumption that the expectations exist.

For example if X_n is the n^{th} trial in the Bernoulli process, $E(X_n) = 1 \cdot p + 0 \cdot q = p$. The expected or average value of the n^{th} toss is p. Needless to say X_n won't ever take the value p (except in the trivial cases p = 0,1). Intuitively, however, if we perform a large number of trials and then average the results, we will get roughly p .

Before we go on to other computations, we need the following important result:

Basic Fact. For any two integer random variables,

$$E(X+Y) = E(X) + E(Y).$$

The surprising thing about this fact is that it holds regardless of whether X and Y are independent or not.

Proof. Let $c_{n_1,n_2} = P((X=n_1)\wedge(Y=n_2))$ be the joint distribution of X and Y. Now X+Y is a new R.V. What is its distribution? Well, we must consider all possible ways that X+Y can take on a given value: $q_k = P(X+Y=k) =$

$$\sum_{\{n_1,n_2:n_1+n_2=k\}} P((X=n_1)\wedge(Y=n_2)) = \sum_{n_1+n_2=k} c_{n_1,n_2} \; .$$

Therefore the expectation of X+Y is:

$$\begin{aligned} E(X+Y) &= \sum_k kq_k = \sum_k k \sum_{n_1+n_2=k} c_{n_1,n_2} \\ &= \sum_{n_1} \sum_{n_2} (n_1+n_2) c_{n_1,n_2} \\ &= \sum_{n_1} \sum_{n_2} n_1 c_{n_1,n_2} + \sum_{n_1} \sum_{n_2} n_2 c_{n_1,n_2} \\ &= \sum_{n_1} n_1 P(X=n_1) + \sum_{n_2} n_2 P(Y=n_2) \\ &= E(X) + E(Y) \; . \end{aligned}$$

This completes the proof.

Notice that it should be intuitively obvious that the expectation of X+Y is $\sum_{n_1} \sum_{n_2} (n_1+n_2)c_{n_1,n_2}$ so the Basic Fact is actually easier than the size of our proof suggests. Namely using probabilistic reasoning we proceed as follows. X+Y takes the "value" n_1+n_2 with probability c_{n_1,n_2}. Adding up all cases gives the expectation:

$$E(X+Y) = \sum_{n_1} \sum_{n_2} (n_1+n_2)c_{n_1,n_2} .$$

Now split the sum and take marginals. The result is E(X) + E(Y).

Let's compute some expectations for the Bernoulli process. First we compute the "hard way" directly from the definition, then we compute using the Basic Fact.

Consider S_n, the number of successes in the first n trials. The distribution for S_n is $p_k = \binom{n}{k}p^k q^{n-k}$. So $E(S_n) = \sum_k k\ p_k = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k}$. Unfortunately we cannot simplify this very easily. On the other hand, $S_n = X_1+X_2+\ldots+X_n$. Hence, $E(S_n) = E(X_1) + E(X_2)+\ldots+E(X_n) = np$, since all of these have the same expectation: p. In addition we have shown that

$$E(S_n) = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = n\ p\ ,$$

a fact that is not so easy to prove.

Next consider the waiting time for the k^{th} head, W_k. The distribution for W_k is $p_n = \binom{n-1}{k-1}p^k q^{n-k}$. Therefore $E(W_k) = \sum_n n\ p_n = \sum_{n=k}^{\infty} n\binom{n-1}{k-1}p^k q^{n-k}$. Again there is no easy way to compute this infinite sum. However, $W_k = T_1+T_2+\ldots+T_k$. Hence, $E(W_k) = E(T_1)+\ldots+E(T_k)$. But all the $T_1,\ldots,T_k$ are equidistributed so in particular they all have the same expectation. Therefore $E(W_k) = k\ E(T_1)$, and we need only compute one expectation: the expectation of $T_1 = W_1$. We shall have to resort to some trickery, but it still isn't too difficult.

$$E(T_1) = \sum_{n=1}^{\infty} n\ P(T_1=n) = \sum_{n=1}^{\infty} n\ q^{n-1}p = p\sum_{n=1}^{\infty} \frac{d}{dq}(q^n) = p\frac{d}{dq}(\sum_{n=1}^{\infty} q^n)$$

$$= p\ \frac{d}{dq}(\frac{q}{1-q}) = p(\frac{1}{(1-q)^2}) = p\cdot\frac{1}{p^2} = \frac{1}{p}\ .$$

Intuitively, it is quite reasonable that $E(T_1) = 1/p$ for if p is large we don't expect to wait very long for a success, while if p is small we expect to wait a long time. As before we get the added bonus of an identity:

$$E(W_k) = \sum_{n=k}^{\infty} n\binom{n-1}{k-1} p^k q^{n-k} = k/p$$

a fact that is quite hard to believe otherwise.

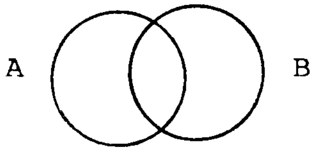
F.* The Inclusion-Exclusion Principle

Imagine that we have a well shuffled deck of cards and that we turn the cards over one at a time. While doing this we call out the names of the cards in their unshuffled order (as in Bridge), beginning with the deuce of clubs and ending with the ace of spades. What is the probability that none of the cards turned over match the name called out when it is turned over? The answer (to an accuracy of 10^{-15}) is $\frac{1}{e}$. This is strange for two reasons: it depends on the number e which shouldn't appear in a finite counting problem, and it doesn't depend on the number of cards in the deck.

We shall prove this result and several others by an important formula called the inclusion-exclusion principle. The proof of this principle will follow easily from the formalism of random variables.

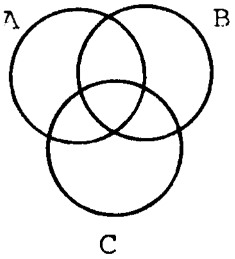
The abstract setting for the principle is the computation of the probability of the union of events in terms of the probabilities of the events and of their intersections. For example if we have two events A and B, then we know that

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



If we refer to the diagram it is clear what this means: $P(A) + P(B)$ "counts" $P(A\cap B)$ twice. Thus we "include" $P(A)$ and $P(B)$ and then "exclude" $P(A\cap B)$.

For three events A,B and C we must include, exclude and then include once again:



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

It is quite easy to think through the proof of this directly. However for the general case it will take a bit more work. Here is the general formula:

$$\boxed{P(A_1\cup A_2\cup\ldots\cup A_n) = \sum_i P(A_i) - \sum_{i<j} P(A_i\cap A_j) + \sum_{i<j<k} P(A_i\cap A_j\cap A_k) - \ldots + \ldots + (-1)^{n+1}P(A_1\cap A_2\cap\ldots\cap A_n)}$$

The Inclusion-Exclusion Principle

Note that the second sum is really a double sum over both i and j but subject to $i<j$, the third is a triple sum and so on.

To prove this principle we introduce a special kind of integer random variable called an <u>indicator</u>. Let A be an event, the <u>indicator</u> of A is the integer random variable I_A corresponding to the question "Did A happen?" More precisely for any sample point $\omega\varepsilon\Omega$,

$$I_A(\omega) = \begin{Bmatrix} 1 & \text{if} & \omega\varepsilon\Lambda \\ 0 & \text{if} & \omega\not\varepsilon\Lambda \end{Bmatrix}$$

One sometimes also sees the notation χ_A for the indicator. We have already encountered such a random variable before. In the Bernoulli process, H_n is the event "the n^{th} toss is heads" and its indicator I_{H_n} is the random variable X_n .

The probability distribution of the indicator I_A is

$$p_n = \begin{cases} P(A^c) & \text{for } n = 0 \\ P(\Lambda) & \text{for } n = 1 \\ 0 & \text{otherwise} \end{cases}$$

Therefore the expectation of I_A is $E(I_A) = 0\cdot P(\Lambda^c) + 1\cdot P(A) = P(A)$. As a result of this we see that all probabilities may be reduced to the computation of expectations; and one could dispense with sample spaces and events altogether and develop probability theory using only random variables and expectations.

We now consider what happens when we add and multiply indicators. The easy operation is multiplication: $I_A I_B = I_{A\cap B}$ should be obvious. Addition, however, is not so easy because the sum of indicators need not be an indicator: $I_A + I_B$ takes value 2 on $A\cap B$. However if we put in a correction term we get an identity: $I_A + I_B = I_{A\cup B} + I_{A\cap B}$. So while multiplication corresponds to intersection, addition does not quite correspond to union.

The last operation we consider is complementation. Here the result is clear: $I_{A^c} = 1 - I_A$. This suggests what we should do in general to compute $I_{A_1\cup A_2\cup \ldots \cup A_n}$ in terms of the A_i's: convert to an intersection by using the DeMorgan law. Thus:

$$\begin{aligned} I_{A_1\cup A_2\cup \ldots \cup A_n} &= 1 - I_{A_1^c\cap A_2^c\cap \ldots \cap A_n^c} \\ &= 1 - I_{A_1^c} I_{A_2^c} \cdots I_{A_n^c} \\ &= 1 - (1-I_{A_1})(1-I_{A_2})\ldots(1-I_{A_n}) \end{aligned}$$

We now multiply out the last expression as in high school algebra:

$$= 1 - [1-\sum_i I_{A_i} + \sum_{i<j} I_{A_i} I_{A_j} - \ldots]$$

$$= \sum_i I_{A_i} - \sum_{i<j} I_{A_i} I_{A_j} + \ldots + (-1)^{n+1} I_{A_1} I_{A_2} \ldots I_{A_n}$$

$$= \sum_i I_{A_i} - \sum_{i<j} I_{A_i \cap A_j} + \ldots + (-1)^{n+1} I_{A_1 \cap A_2 \ldots \cap A_n}$$

Finally we take the expectation of this expression using the Basic Fact of expectations. The result is the inclusion-exclusion principle.

We now return to our first question. Think of the situation as follows. Start with a new unshuffled deck and then shuffle it. The result is a random permutation of the unshuffled deck. What is the probability that no card is in the same position in both the unshuffled and the shuffled decks?

To be more precise consider the integers 1,...,n instead of the 52 cards. The sample space is the set Ω of all permutations of 1,...,n. Thus $|\Omega| = n!$ The notation for permutations is $\binom{1\ 2\ 3\ \ldots n}{i_1 i_2 i_3 \ldots i_n}$, where one should think of the top row as the unshuffled integers and the lower row as the shuffled ones. A <u>fixpoint</u> of a permutation is a number j so that $i_j = j$, i.e. the same number j appears twice in one column in our notation. For example let n = 3. There are 6 permutations with number of fixpoints as follows:

permutation	number of fixpoints
$\binom{123}{123}$	3
$\binom{123}{213}$	1
$\binom{123}{231}$	0
$\binom{123}{321}$	1
$\binom{123}{312}$	0
$\binom{123}{132}$	1

Let F be the event "there is at least one fixpoint". We want to compute $P(F^c)$. Counting F directly is not very easy, but we can write F as the union of events that we can count quite easily. Let A_i be the event "i is a fixpoint". Then $F = A_1 \cup A_2 \cup \ldots \cup A_n$. Since the A_i are not disjoint we must apply the principle of inclusion-exclusion:

$$P(F) = \sum_i P(A_i) - \sum_{i<j} P(A_i \wedge A_j) + \ldots$$

Now an element of A_1 has 1 as a fixpoint so it is just a permutation of $\{2,\ldots,n\}$. Therefore $|A_1| = (n-1)!$ and similarly $|A_i| = (n-1)!$ Any element of $A_1 \wedge A_2$ has two fixpoints so it is a permutation of $\{3,\ldots,n\}$. So $|A_1 \wedge A_2| = (n-2)!$ More generally $|A_1 \wedge A_2 \wedge \ldots \wedge A_k| = (n-k)!$ If we divide by n! we get the probabilities, e.g. $P(A_1 \wedge \ldots \wedge A_k) = \frac{(n-k)!}{n!} = (n)_k^{-1}$. Substituting these into our formula for P(F) gives us:

$$P(F) = \binom{n}{1} \frac{1}{(n)_1} - \binom{n}{2} \frac{1}{(n)_2} + \binom{n}{3} \frac{1}{(n)_3} - \ldots$$

$$= \frac{(n)_1}{1!} \frac{1}{(n)_1} - \frac{(n)_2}{2!} \frac{1}{(n)_2} + \frac{(n)_3}{3!} \frac{1}{(n)_3} - \ldots$$

$$= \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \ldots + (-1)^{n+1} \frac{1}{n!}$$

From calculus you should immediately recognize this expression as the beginning of the expansion for $1-e^x$ when $x = -1$. This expansion converges so extremely rapidly that it is essentially $1-e^{-1}$ when, say, n is larger than 7. We conclude that

$$P(F^c) = \frac{1}{e} \text{ to high accuracy (when } n>7).$$

We consider another application. Suppose we have an infinite collection of balls labelled 1,2,3,..., and suppose we have n boxes. If we drop the balls into the boxes sequentially, how long do we have to wait until every box contains at least one ball? If this sounds devoid of physical interest consider the following mathematically equivalent statement. Suppose we have a molecular beam firing molecules at a target crystal. Assume that a molecule adheres to the crystal if it strikes an unoccupied lattice site and rebounds (and is lost) if it strikes a previously occupied site. If we assume that the molecules are fired at random at the crystal sites, how long must we wait until all the crystal sites are covered? This problem and perturbations of it are very real problems in surface physics.

The answer to our question will of course be a probability distribution. Let W be the waiting time until all the boxes are occupied. We want to compute $p_k = P(W \leq k)$. This is the probability that if we place k balls into n boxes, then all the boxes are occupied. Let A_i be the event "the i^{th} box is empty". Then $(W \leq k) = A_1^c \cap A_2^c \cap \ldots \cap A_n^c$. By the inclusion-exclusion principle,

$$\begin{aligned} P(W \leq k) &= P(A_1^c \cap A_2^c \cap \ldots \cap A_n^c) \\ &= 1 - P(A_1 \cup A_2 \cup \ldots \cup A_n) \\ &= 1 - \sum_i P(A_i) + \sum_{i<j} P(A_i \cap A_j) - \ldots \end{aligned}$$

Now the sample space Ω consists of all placements (Maxwell-Boltzmann) of k balls into n boxes. Thus $|\Omega| = n^k$. The event A_1 consists of all placements of k balls into the last n-1 boxes. Thus $|A_1| = (n-1)^k$. Similarly $|A_1 \cap A_2| = (n-2)^k$ and so on. So the probability of A_i is $P(A_i) = \frac{(n-1)^k}{n^k} = (1-\frac{1}{n})^k$, that of $A_i \cap A_j$ is $P(A_i \cap A_j) = \frac{(n-2)^k}{n^k} = (1-\frac{2}{n})^k$, and so on. Therefore:

$$P(W\leq k) = 1 - \binom{n}{1}(1-\tfrac{1}{n})^k + \binom{n}{2}(1-\tfrac{2}{n})^k - \dots .$$

As a final application of these ideas, we consider the problem of writing max $(x_1,x_2,\dots,x_n)$, for n real numbers $x_1,\dots,x_n$, in terms of their minima. We won't go through all the details. The idea is to consider the set of real numbers as being the sample space Ω and to use the indicators $I_{(-\infty,x)}$. For example, $I_{(-\infty,x)}I_{(-\infty,y)} = I_{(-\infty,x)\cap(-\infty,y)} = I_{(-\infty,\min(x,y))}$ and $I_{(-\infty,x)} + I_{(-\infty,y)} = I_{(-\infty,\max(x,y))} + I_{(-\infty,\min(x,y))}$. We leave it as an exercise to show that:

$$\max(x_1,x_2,\dots,x_n) = \sum_i x_i - \sum_{i<j}\min(x_i,x_j) + \sum_{i<j<k}\min(x_i,x_j,x_k)$$

$$- \dots + \dots + (-1)^{n+1}\min(x_1,x_2,\dots,x_n).$$

2. General Random Variables

So far we have considered only integer random variables. We now allow random variables to take any real value. Unfortunately technical difficulties will appear that didn't occur with integer random variables. We begin with an example so that we can gradually work our way into the difficulties.

Consider the process of dropping a point on the interval [0,a]. Intuitively the point is just as likely to fall on one part of [0,a] as another. For example it should be just as probable for the point to fall on the left half of the interval as to fall on the right half. More generally, the probability that the point falls in any given subinterval is proportional to the length of that subinterval. Unfortunately this leads to the inescapable conclusion that the probability of the point taking any one particular value x is <u>zero</u>.

So we see that the intuitive concept of an integer random variable, i.e. of a variable which takes its values with certain probabilities, is inadequate for describing the phenomenon of a general random variable. In fact there is the an intriguing philosophical paradox here: how can the point land anywhere at all if the probability of its landing in any one place is zero? We will avoid such seeming paradoxes by decreeing that the probabilistic structure of a random variable is given by the probabilities that it takes values in <u>intervals</u>. More precisely if X is a random variable, the probabilistic structure of X is given by the probability that X is between c and d for any real numbers $c \leq d$. We write $P(c \leq X \leq d)$ for this probability. For example, if X is the random variable corresponding to a point dropped at random on $[0,a]$, then for any pair of real numbers $c<d$ in $[0,a]$,

$$P(c \leq X \leq d) = \frac{d-c}{a} .$$

As another example, let X be an integer R.V. Then

$$P(c \leq X \leq d) = \sum_{c \leq n \leq d} p_n .$$

There is a neat way to express the probabilistic structure of random variables in general: the <u>(probability) distribution function</u>. We define this to be the function

$$F(x) = P(X \leq x) .$$

To compute probabilities on "half-open" intervals we use the fact that:

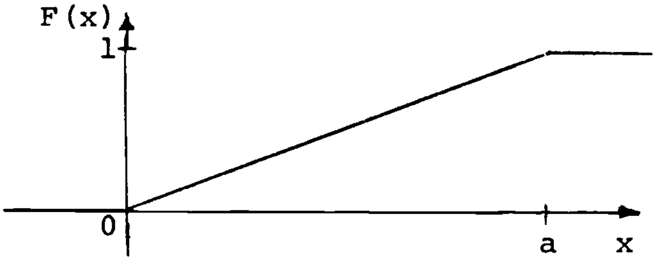
$$P(c<X\leq d) = P(X\leq d) - P(X\leq c) = F(d) - F(c) \ .$$

For intervals in general we use limits and the above formula. Therefore the probabilistic structure of a random variable is completely determined by its distribution function.

Consider once again the random variable X corresponding to dropping a point at random on [0,a]. The distribution function of X is

$$F(x) = P(X\leq x) = \begin{cases} 0 & \text{if } x<0 \\ x/a & \text{if } 0\leq x\leq a \\ 1 & \text{if } x>a \end{cases}$$

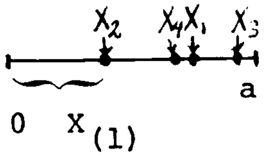
When a random variable has this distribution function we shall say that it is uniformly distributed on [0,a]. Typically, distribution functions will have "kinks".



Graph of the distribution function of a uniformly distributed random variable

We see that the probabilistic meaning of dropping a point at random is that we have a random variable X uniformly distributed on $[0,a]$. We might also say that we are "choosing" or "sampling" a point at random from $[0,a]$. The process of sampling a sequence of n points at random from $[0,a]$ is called the Uniform process. More precisely, a Uniform process of sampling n points from $[0,a]$ is a sequence of n independent random variables $X_1, X_2, \ldots, X_n$ uniformly distributed on $[0,a]$. It is the continuous analog of the finite sampling process in chapter II. Note that we do not have to distinguish between sampling with or without replacement because the probability that any two of the sampled points coincide is zero.

A typical question one may ask about the Uniform process is: what is the length of the gap between zero and the smallest point of the n dropped points. The naive answer is "it depends on which X_i is the smallest". We shall answer the question with a probability distribution function. More precisely write $X_{(1)}$ (pronounced "X order 1") for the smallest point: $X_{(1)} = \min(X_1, \ldots, X_n)$. Then the probabilistic answer to our question is the distribution function of $X_{(1)}$. To compute this note that the event $(X_{(1)} > t)$ is the same as saying that all the X_i are greater than t. Hence:



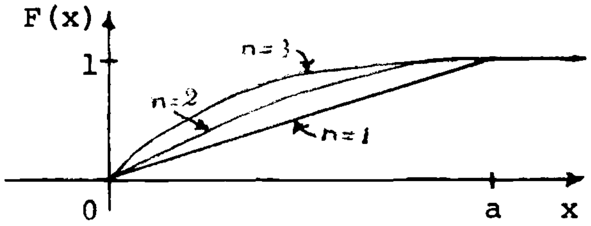
$$P(X_{(1)} > t) = P((X_1 > t) \wedge (X_2 > t) \wedge \ldots \wedge (X_n > t))$$

$$= P(X_1 > t) P(X_2 > t) \ldots P(X_n > t)$$

$$= \left(\frac{a-t}{a}\right)^n$$

[Of course to justify this computation rigorously we must define independence for arbitrary random variables. We will do this in the next section.] Therefore the distribution function of $X_{(1)}$ is

$$F_{(1)}(x) = P(X_{(1)} \leq x) = 1-P(X_{(1)} > x) = 1-(\frac{a-x}{a})^n .$$



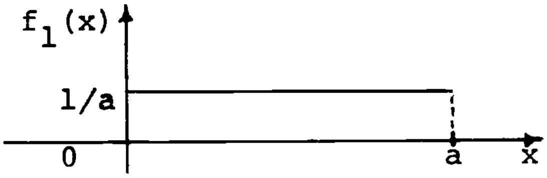
Graph of the distribution function of $X_{(1)}$

The distribution function is more and more "concentrated" near 0 as n increases: the more points one drops, the more likely that the first gap is small.

We need a way to express more clearly the fact that the distribution is more concentrated near 0 for larger n. Indeed as we shall see, the distribution of a R.V. is not a very good way to visualize the behavior of the R.V. A better way is to use the density of the R.V.

The density of a R.V. X is the derivative (if it exists) of the distribution: $f(x) = \frac{d}{dx} F(x)$. By calculus, $\int_{-\infty}^{x} f(u)du = F(x)$. For example the density of X_1 in the uniform process is

$$f_1(x) = \begin{cases} 1/a & 0<x<a \\ 0 & x<0 \text{ or } x>a \end{cases}$$

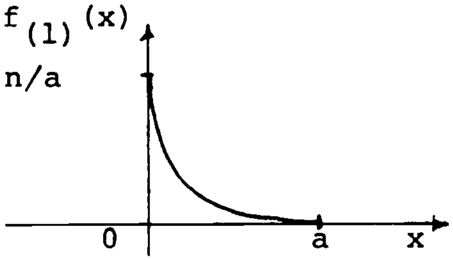


Graph of the density of X_1

Using density we see much more clearly why X_1 is said to be uniformly distributed on [0,a]: its density is constant on [0,a],

On the other hand, the density of $X_{(1)}$ is

$$f_{(1)}(x) = \begin{cases} \frac{n(a-x)^{n-1}}{a^n} & 0<x<a \\ 0 & x<0 \text{ or } x>a \end{cases}$$



Graph of the density of $X_{(1)}$.

Notice how the density is sharply peaked at x=0 just as we intuitively would expect.

The Concept of Random Variable

We are now ready to give rigorous definitions of the intuitive ideas in the last section.

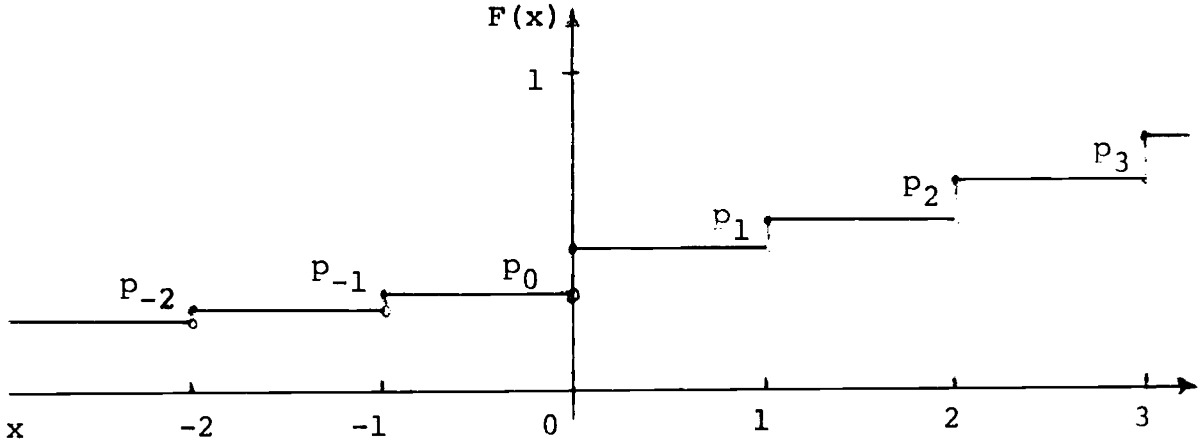
Definition. A random variable X is a function from a sample space Ω to the real numbers, with the property that the subsets $(X\leq x) = \{\omega\varepsilon\Omega: X(\omega)\leq x\}$ are events of Ω for all real numbers x. The (probability) distribution function of a random variable X is the function

$$F(x) = P(X\leq x).$$

As similarly noted for integer R.V.'s, the technical assumption that the subsets $(X\leq x)$ are events will never bother us. We state it for purely grammatical reasons.

Integer Random Variables

Integer R.V.'s are characterized by the fact that their distribution functions are constant except at integers, where they have discontinuous jumps.



Graph of the distribution function of an integer random variable.

Being a discontinuous function, the distribution function of an integer R.V. is rather unpleasant to deal with. As a result one generally considers instead the probability distribution $p_n = P(X=n)$. It is unfortunate that F(x) and p_n are both referred to as the distribution of an integer R.V.

Continuous Random Variables

A random variable X is a continuous random variable if its distribution function F(x) is continuous and piecewise differentiable. The derivative $f(x) = F'(x)$ is called the

<u>density</u> of X. It is the continuous analogue of the probability distribution p_n of an integer R.V. This can be made quite precise using infinitesimals: the probability that X takes a given value x is the infinitesimal f(x)dx. In other words, the probability that X takes a value in a very small interval [x,x+h] is close to f(x)h, the smaller the interval, the closer the approximation.

This suggests that the probability for a continuous random variable X to take a given value x is not quite zero but rather infinitesimal, if $f(x)\neq 0$. So although (X=x) is an unlikely event, it is not impossible. We will write **dens(X=x)** for the density f(x) of X at x. However, one should take caution when using this notation: **dens(X=x)** does not act like a probability P(X=x). To give a concrete example, let X be uniformly distributed on [0,1]. Then 2X is uniformly distributed on [0,2]. Hence **dens(X=x)**$=1 \neq \frac{1}{2} =$ **dens(2X=2x)** , even though the events (X=x) and (2X=2x) are obviously the same. In general, before performing any calculations involving densities, one should first convert them to probabilities. For example,

$$\mathbf{dens(X=x)} = \frac{d}{dx} P(X\leq x) = \frac{d}{dx} P(2X\leq 2x)$$

$$\mathbf{dens(2X=2x)} = \frac{d}{d(2x)} P(2X\leq 2x) = \frac{1}{2}\frac{d}{dx} P(2X\leq 2x)$$

therefore, **dens(X=x)** = 2 **dens(2X=2x)**.

The density of X acts precisely as a mass density on the real line, a familiar concept in calculus. Thus, for example, to compute $P(a<X\leq b)$ we must integrate the density:

$$P(a<X\leq b) = \int_a^b f(x)dx.$$

In the case of an integer R.V. we get a sum:

$$P(k\leq X\leq n) = \sum_{i=k}^{n} p_i \ .$$

The integral is the continuous analogue of a sum.

<u>Independence</u>

The concept of the independence of two arbitrary R.V.'s ought to be obvious, given the definition in the integer case. Namely two R.V.'s X and Y are <u>independent</u> if the events $(X\leq x)$ and $(Y\leq y)$ are independent for any pair of real numbers x and y:

$$P((X\leq x)\wedge(Y\leq y)) = P(X\leq x)P(Y\leq y) \ .$$

Properties of Densities and Distributions

The distribution function F(x) of an arbitrary R.V. satisfies:

(1) $F(x) \leq F(y) \quad \text{if } x \leq y$

(2) $\lim\limits_{x \to -\infty} F(x) = 0$

(3) $\lim\limits_{x \to \infty} F(x) = 1$

(4) F is left continuous, i.e. $\lim\limits_{\substack{y \to x \\ y > x}} F(y) = F(x)$

All these are obvious consequences of the definition of the distribution function. It is an interesting exercise to show the converse: any function F(x) satisfying (1)-(4) is the distribution function of some R.V. X on some sample space.

When X is a continuous R.V., its density f(x) satisfies properties analogous to those of the distribution p_n of an integer R.V. Namely,

(1) $f(x) \geq 0$

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$

Joint Distribution and Joint Density

Just as we did for integer random variables, we measure the correlation of two arbitrary R.V.'s by using a joint distribution function. The joint distribution function of R.V.'s X and Y is

$$F(x,y) = P((X \leq x) \wedge (Y \leq y)).$$

If X and Y are continuous, then they also have a joint density:

$$\text{dens}(X=x,\ Y=y) =$$
$$f(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y} F(x,y) \ .$$

In terms of infinitesimals, the probability that X takes the value x and Y takes the value y is f(x,y) dx dy. As with ordinary densities, be careful not to treat dens(X=x, Y=y) as a probability.

Suppose that F_X, F_Y and f_X, f_Y denote the distribution functions and densities of the continuous R.V.'s X and Y respectively. We can recover these from their joint counterparts:

$$\int_{-\infty}^{\infty} f(x,y)\,dy = f_X(x)$$
$$\int_{-\infty}^{\infty} f(x,y)\,dx = f_Y(y)$$

We call these the marginal densities or marginals.

$$F_X(x) = \lim_{y\to\infty} F(x,y)$$
$$F_Y(y) = \lim_{x\to\infty} F(x,y)$$

We won't have much use for the last two formulas.

In terms of the joint distribution and joint density, two random variables X,Y are independent if and only if

$$F(x,y) = F_X(x)\ F_Y(y)$$

or $$f(x,y) = f_X(x)\ f_Y(y)$$

Expectation

For an integer R.V. X, the expectation of X is the mean or average value of X: $E(X) = \sum_n np_n$. For a continuous R.V.X, the *expectation* is the continuous analogue: $E(X) = \int_{\infty}^{\infty} xf(x)dx$, if it exists. One should immediately recognize this as the *center of mass* of the mass density given by f(x).

The expectation of a continuous R.V. also satisfies the property we found so useful for integer R.V.'s:

Basic Fact. For any two continuous random variables X *and* Y ,

$$E(X+Y) = E(X) + E(Y).$$

Proof. This is essentially the same proof as in the integer case.

$$E(X+Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dx\,dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\ f(x,y)dx\ dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)dx\ dy$$

$$= \int_{-\infty}^{\infty} x\ f_X(x)dx + \int_{-\infty}^{\infty} yf_Y(y)dy$$

$$= E(X) + E(Y).$$

3. The Uniform Process

We now make a detailed investigation of this process in order to illustrate the concepts we have just introduced. Recall that the Uniform process of sampling n points from

$[0,a]$ is the same as a sequence of independent random variables $X_1, X_2, \ldots, X_n$ each being uniformly distributed on the interval $[0,a]$. For example, these random variables might be the measurements of the heights of a random sample of n people. If we wish to ignore the order in which the people are measured, we simply write down the heights in increasing order. We call this new sequence the order statistics of the original sample. In effect we "forget" what the order of sampling was and consider only the set of n measurements.

To be more precise we introduce the following notation:

$$X_{(1)} = \min (X_1, \ldots, X_n)$$

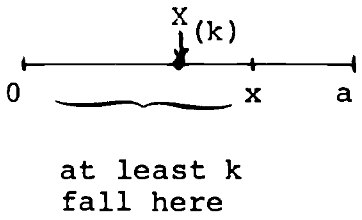
$$X_{(2)} = \text{next larger point after } X_{(1)}$$

$$\ldots$$

$$X_{(n)} = \max (X_1, \ldots, X_n)$$

How are the order statistics distributed? What are their joint distributions? What are the distributions of the gaps between successive order statistics? Unlike the gaps in the Bernoulli process, these are not independent; for if one is big, the others must be small. What are the joint distributions of the gaps? We shall now answer these and other questions.

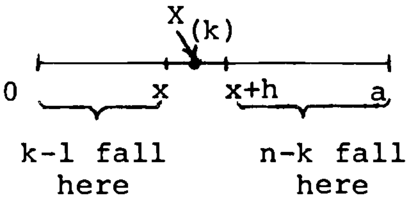
Let $F_{(k)}(x)$ and $f_{(k)}(x)$ denote the distribution and density of the k^{th} order statistic $X_{(k)}$. Thus $F_{(k)}(x) = P(X_{(k)} \leq x)$. Now $(X_{(k)} \leq x)$ is the event "at least k of the n points fall in the interval [0,x]". We decompose this event according to the number of points that actually fall in [0,x]. Therefore:



$$F_{(k)}(x) = P(X_{(k)} \leq x) = \binom{n}{k}\left(\frac{x}{a}\right)^{k}\left(\frac{a-x}{a}\right)^{n-k} + \binom{n}{k+1}\left(\frac{x}{a}\right)^{k+1}\left(\frac{a-x}{a}\right)^{n-k-1} +$$

$$\ldots + \binom{n}{n}\left(\frac{x}{a}\right)^{n}.$$ For example, the first summand is the probability that exactly k points fall in [0,x], and hence exactly n-k fall in (x,a]. Similarly for the other summands. Needless to say this expression is awkward.

Consider now the density $f_{(k)}(x)$. We first compute that



$P((x<X_{(k)} \leq x+h)$ and no other $X_{(j)}$ falls in $[x,x+h]) =$

$$\binom{n-1}{k-1}\cdot\left(\frac{x}{a}\right)^{k-1}\cdot n\cdot\frac{h}{a}\cdot\left(\frac{a-x-h}{a}\right)^{n-k} .$$

Here one of the n points falls in [x,x+h]: probability $n\cdot\frac{h}{a}$.

Next, k-1 of the remaining n-1 points fall in [0,x]: probability $\binom{n-1}{k-1}(\frac{x}{a})^{k-1}$. Finally the remaining n-k points fall in [x+h,a]: probability $(\frac{a-x-h}{a})^{n-k}$. Unfortunately what we really want is $P(x<X_{(k)} \leq x+h)$. This appears to be a much more complicated computation.

However, we never really have to compute this expression for the following reason. If more than one of the $X_{(j)}$ fall in [x,x+h], the resulting probability involves a factor of $(\frac{h}{a})^2$ (or possibly even a higher power of $\frac{h}{a}$). Thus

$$P(x<X_{(k)} \leq x+h) = n\binom{n-1}{k-1} \frac{x^{k-1}\cdot h\cdot (a-x-h)^{n-k}}{a^n} + \frac{h^2}{a^2} \cdot \text{(complicated expression)}$$

Now divide by h and take the limit as h→0:

$$f_{(k)}(x) = \lim_{h\to 0} \frac{P(x<X_{(k)} \leq x+h)}{h} = \lim_{h\to 0} n\binom{n-1}{k-1}\frac{x^{k-1}(a-x-h)^{n-k}}{a^n} + \frac{h}{a} \cdot \text{(crud}$$

$$f_{(k)}(x) = \left\{ \begin{array}{ll} n\binom{n-1}{k-1} \dfrac{x^{k-1}(a-x)^{n-k}}{a^n} & \text{, if } 0\leq x\leq a \\ 0 & \text{otherwise} \end{array} \right\}$$

We never have to compute the complicated expression because no matter what it is, it disappears when we let h go to zero. We shall use this trick repeatedly. In fact it is precisely because we can make this kind of simplification that the density is so much more computable than the distribution.

We now mention an interesting application. The function $f_{(k)}(x)$ is a probability density so it integrates to 1:

$$1 = \int_{-\infty}^{\infty} f_{(k)}(x)dx = \int_0^a f_{(k)}(x)dx = \int_0^a n\binom{n-1}{k-1}\frac{x^{k-1}(a-x)^{n-k}}{a^n}\,dx.$$

Therefore:

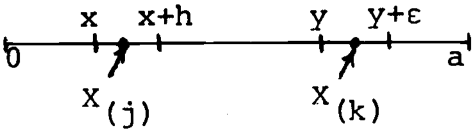
$$\int_0^a x^{k-1}(a-x)^{n-k}dx = \frac{a^n}{n\binom{n-1}{k-1}}\ .$$ Thus just as integer R.V.'s allow us to compute certain infinite series probabilistically, continuous R.V.'s furnish a technique for computing certain definite integrals. We shall see more of these as we go on.

Next we consider the joint distribution of two order statistics. For example, how does the tenth point influence the twentieth? This is an important question in biostatistics, because of the necessity of biologists to rely on small samples. Tables of order statistics allow one to detect deviations from randomness in a relatively small sample.

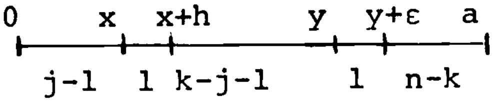
As with the above computation it is much more convenient to compute the joint density. Let $X_{(j)}$, $X_{(k)}$ be two of the order statistics, $j<k$. Then the joint distribution is

$$F_{(j,k)}(x,y) = P((X_{(j)} \le x) \cap (X_{(k)} \le y))$$

and the density is $f_{(j,k)}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y} F_{(j,k)}(x,y)$. Again as with the computation above we need only compute the probability of the event "$X_{(j)}$ falls in $[x,x+h]$, $X_{(k)}$ falls in $[y,y+\varepsilon]$ and no other points fall in these intervals". We think of these two intervals as dividing $[0,a]$ into 5 boxes into which we drop n distinguishable balls with occupation numbers: $j-1$, 1, $k-j$, 1, $n-k$.



There are $\binom{n}{j-1,1,k-j-1,1,n-k}$ ways to place the n balls with these occupation numbers. Therefore the event in question has probability



Boxes and occupation numbers

$$\binom{n}{j-1,1,k-j-1,1,n-k} \left(\frac{x}{a}\right)^{j-1} \frac{h}{a} \left(\frac{y-x-h}{a}\right)^{k-j-1} \frac{\varepsilon}{a} \left(\frac{a-y-\varepsilon}{a}\right)^{n-k} .$$

Dividing by $h\varepsilon$ and letting $h \to 0$ and $\varepsilon \to 0$ gives the joint density:

$$f_{(j,k)}(x,y) = \binom{n}{j-1,1,k-j-1,1,n-k} \frac{x^{j-1}(y-x)^{k-j-1}(a-y)^{n-k}}{a^n}$$

or

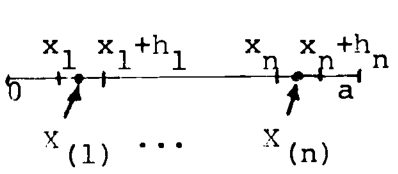
$$f_{(j,k)}(x,y) = \begin{cases} \dfrac{n!}{(j-1)!(k-j-1)!(n-k)!} \dfrac{x^{j-1}(y-x)^{k-j-1}(a-y)^{n-k}}{a^n} & \text{if } x<y \\ 0 & \text{if } x \geq y \end{cases}$$

Finally we consider the joint density of all n order statistics. Let $x_1<x_2<\ldots<x_n$ be real numbers in $[0,a]$. Now

$$P((x_1<X_{(1)} \leq x_1+h_1) \cap (x_2<X_{(2)} \leq x_2+h_2) \cap \ldots \cap (x_n<X_{(n)} \leq x_n+h_n))$$

$$= n! \ \frac{h_1}{a} \cdot \frac{h_2}{a} \cdot \ldots \cdot \frac{h_n}{a} \text{ , because}$$

there are n! ways of placing the n points in the intervals $[x_1,x_1+h_1],\ldots,[x_n,x_n+h_n]$. The h_i's are chosen so small that there is no overlap. Therefore the joint density of all n order statistics is:



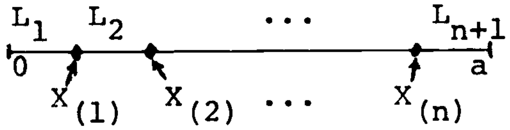
$$f(x_1,\ldots,x_n) = \begin{cases} \dfrac{n!}{a^n} & \text{if } 0\le x_1<x_2<\ldots<x_n\le a \\ 0 & \text{otherwise} \end{cases}$$

Like all densities, $f(x_1,\ldots,x_n)$ integrates to 1. Thus we get the interesting multiple integral:

$$\int_0^a \left(\int_0^{x_n} \cdots \left(\int_0^{x_2} dx_1\right)dx_2\right)\ldots\Bigg)dx_n = \frac{a^n}{n!}$$

This is reasonable because the conditions $0\le x_1<x_2\ldots<x_n\le a$ determine a "pyramid" cut off the n-cube of side a at one corner.

The gaps of the uniform process are the distances between successive points in increasing order. The gap between 0 and $X_{(1)}$ is written L_1, the gap between $X_{(j)}$ and

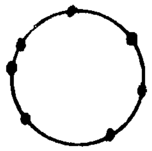


$X_{(j+1)}$ is written L_{j+1}, and the gap between $X_{(n)}$ and a is L_{n+1}. The order statistics may be written in terms of the gaps:

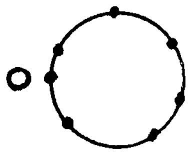
$$X_{(k)} = L_1 + L_2 + \dots + L_k \ .$$

The gaps are not independent: if one is large the others must be small. But the gaps are nevertheless equidistributed! When we have conditional probabilities, we will be able to prove this rigorously. However one can prove this probabilistically. Since one of our main objectives is to learn to think probabilistically, this kind of proof is actually preferable.

Imagine that we drop n+1 points on a circle of circumference a. It is intuitively obvious by symmetry that the gaps (measured along the circumference) so obtained are all equidistributed. On the other hand, this experiment is statistically equivalent to the following experiment. Fix one point (call it O) on the circle and then drop n more points at random. If we cut the circle at O and stretch it out



7 points at random
on a circle



6 points at random
on a circle plus a
fixed point O

over the interval [0,a], then the gap distributions on [0,a] are the same as those on the circle (the probability that another of the n points falls at the same place as O is zero). Therefore the gap distributions on [0,a] are all equidistributed. This completes the proof.

Therefore all the gaps are distributed the same as $L_1 = X_{(1)}$. We already computed the density of $X_{(1)}$ so the density of any gap L_i is

$$f(x) = \frac{n}{a^n} (a-x)^{n-1} \quad \text{on } [0,a].$$

The expectation of L_i is given by:

$$E(L_i) = \int_0^a x \frac{n}{a^n}(a-x)^{n-1}dx = \frac{n}{a^n} \int_0^a x(a-x)^{n-1}dx ,$$

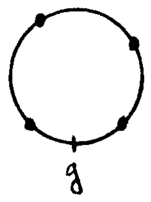
but there is an easier way to compute this. Since

$$E(L_1) = E(L_2)=\ldots= E(L_{n+1}) ,$$

and since $L_1 + L_2+\ldots+ L_{n+1} = a$, we conclude that $E(L_i) = \frac{a}{n+1}$ by the Basic Fact. We can now appreciate the power of the Basic Fact, for the L_i's are not independent.

Similarly we can compute $E(X_{(i)})$ quite easily. For $X_{(i)} = L_1+\ldots+L_i$ implies that $E(X_{(i)}) = E(L_1)+\ldots+E(L_i) = \frac{i \cdot a}{n+1}$. This is certainly what one would intuitively expect, but a direct computation would be tedious.

Consider now a seeming paradox. Suppose we label a reference point g on a circle of circumference a, then we drop n points at random. What is the expected length of the gap that includes the reference point g? The answer is $\frac{2a}{n+1}$, not $\frac{a}{n}$ as one might intuitively expect. The paradox lies not in any contradiction but rather in having a false intuition. Think of the experiment in reverse order: drop n points at random and then choose a reference point g . Then g is more likely to fall in a longer gap simply because it is longer. The seeming paradox comes from the impression that one is performing the following quite different experiment: drop n points at random on a circle and then pick a gap at random (i.e. any gap is as likely to be chosen as any other). This experiment does indeed have expectation $\frac{a}{n}$.



4 points at random and a reference point g.

4. Table of Probability Distributions

Random variables are a central concept in the theory of probability. For example we saw that the uniform process is simply the study of n independent, uniformly distributed random variables. One could regard probability theory abstractly as the study of certain functions on sample spaces, which satisfy certain laws. However this would miss the point, because it is the examples that make the theory, and we can only learn probability theory by carefully studying the examples, especially the important ones.

Random variables are classified by their distributions. And when one speaks of a distribution one usually has a standard model in mind. Learning probability theory therefore requires learning not just the distribution but also the natural phenomena that give rise to them. We will now make a list of distributions and models. We will add to our list in subsequent chapters.

Bernoulli distribution. X has the Bernoulli distribution if X is an integer R.V. which takes just two values, 0 and 1. This distribution depends on one parameter, $p=P(X=0)$. The standard model for this distribution is a toss of a biased coin, X_i, with bias p in the Bernoulli process.

Binomial distribution. X has the binomial distribution if X is an integer R.V. and $P(X=k) = \binom{n}{k}p^k q^{n-k}$. The binomial distribution depends on two parameters n and p. The standard model is S_n, the number of heads in the first n tosses of the Bernoulli process. Here p measures the "bias" of the coin.

Geometric (Pascal) distribution. X has the geometric distribution if X is an integer R.V. and $P(X=n) = q^{n-1}p$. The standard models are the waiting time W_1 for the first head in the Bernoulli process and the gap T_k between the $(k-1)^{st}$ and k^{th} occurrences of heads in the Bernoulli process.

Negative Binomial distribution. X has this distribution if X is an integer R.V. and $P(X=n) = \binom{n-1}{k-1}q^{n-k}p^k$. This distribution has one parameter k . The standard model is the k^{th} waiting time W_k of the Bernoulli process.

Uniform distribution. X has this distribution if X is a continuous R.V. with density $f(x) = \left\{\begin{array}{ll} 1/a & \text{if } 0 \le x \le a \\ 0 & \text{otherwise} \end{array}\right\}$.

The standard model is any X_i, "dropping a point at random", in the Uniform process. Here the parameter a is the length of the interval on which one is dropping (or sampling) points.

Distributions of order statistics. These are sometimes called the *Dirichlet distributions.* X has one of these distributions if X is a continuous R.V. and its density is

$$f(x) = \begin{cases} n\binom{n-1}{k-1} \dfrac{x^{k-1}(a-x)^{n-k}}{a^n} & \text{if } 0 \le x \le a \\ 0 & \text{otherwise.} \end{cases}$$

There are three parameters a,n and k. The standard model is the k^{th} order statistic $X_{(k)}$ of n points dropped at random on [0,a]. The gaps L_i between the order statistics are all models for the distribution having k = 1.

Distribution	type	parameter(s)	model(s)
Bernoulli	integer	p	X_i in the Bernoulli process
Binomial	integer	n,p	S_n (X_i when n=1) in the Bernoulli process
Geometric (Pascal)	integer	p	W_1 or any T_k in the Bernoulli process
Negative Binomial	integer	k	W_k " " "
Uniform	continuous	a	X_i in the Uniform process
Dirichlet	continuous	a,n,k	$X_{(k)}$ (L_i when k=1) in the Uniform process

Distribution	Probability distribution or density	Expectation
Bernoulli	$p_0 = q = 1-p,\ p_1 = p$	p
Binomial	$p_k = \binom{n}{k} p^k q^{n-k}$	np
Geometric	$p_n = q^{n-1}p$	1/p
Negative Binomial	$p_n = \binom{n-1}{k-1} q^{n-k} p^k$	k/p
Uniform	$f(x) = 1/a$ on [0,a]	a/2
Dirichlet	$f(x) = n\binom{n-1}{k-1} x^{k-1}(a-x)^{n-k}a^{-n}$ on [0,a]	$ka/(n+1)$

Table of Bernoulli and Uniform Distributions

5. Exercises for

Chapter III Random Variables

Integer Random Variables

1. The thirteen diamonds are taken from a deck of cards and are thoroughly shuffled. One diamond is drawn at random and scored as follows: two through ten score as their rank, face cards score ten and the ace scores eleven. Let S be the score. Describe the sample space and probability measure used in this problem. Write out S explicitly as a function on the sample space. Write out the probability measure P explicitly as a function. Do S and P have the same domain?

2. In San Francisco, a drunk leaves a bar and every 10 seconds staggers either one yard down the street with probability 3/4 or one yard up the street with probability 1/4. Where is the drunk after one minute? after two minutes? What is his most likely location in each case? How is the most likely location varying in time?

3. A machine that produces screws is subject to occasional surges in its power supply. These occur independently during each second of time with 90% probability and the machine produces one screw every second. In one version of the machine there is a fuse that shuts off the machine permanently when a power surge occurs. We wish to know how many screws the machine produces after it is turned on. Which random variable in the Bernoulli process coresponds to this question? Answer the question.

4. Another version of the machine in exercise 3 has a temporary circuit breaker so that during a power surge the functioning of

the machine is interrupted only for one second. We run the machine for one minute and wish to know how many screws are produced. Which random variable in the Bernoulli process corresponds to this question? Answer the question.

5. In a bridge game the deck is thoroughly shuffled and dealt. You are dealt a hand containing four spades. How many spades was your partner dealt?

6. Three office workers take a coffee break. They choose one of their number at random to pay for the coffee as follows. All three flip a coin simultaneously and the one having a different outcome pays for the coffee. If all coins come up the same, they flip the coins again. How long does it take to determine who pays for the coffee?

7. If one has a coin with a bias $p \neq 1/2$, one can nevertheless use it to synthesize a fair coin by the following trick. Flip the coin twice. If the two tosses come out different, we can say that we got heads if the first toss was heads and tails otherwise. If the two tosses were the same, we toss the coin two more times and proceed as above. Show that this produces a fair coin toss. How many tosses of the biased coin are required to produce one "fair toss"?

8*. Given any bias p between 0 and 1 and a fair coin, one can synthesize a biased coin toss with this bias as follows. Write the binary expansion of $q = 1 - p$. This is just a sequence of zeroes and ones after the decimal point (binary point?). Now start tossing the fair coin. When we get a head write down a

1 and for a tail write a 0. Compare the sequence we obtain with the binary expansion of q. Continue tossing until the first time that the two sequences differ. At this point we stop and record what happened on the last toss. Show that what we record is equivalent to a biased coin toss with bias p. How long does it take to complete such a toss? Does it depend on p?

Independence

9. In exercise 2, is the position of the drunk after one minute independent of his position after two minutes?

10.* Prove that if X and Y are independent random variables and if $f(x)$ and $g(y)$ are two functions, then $f(X)$ and $g(Y)$ are also independent random variables.

Expectation

11. What is the distribution of S in exercise 1? What is its average value?

12. Compute the average position of the drunk in exercise 2 after one minute and after two minutes. How is the drunk's average position changing in time? How do these questions differ from the questions asked in exercise 2?

13. In the game of Chuck-A-Luck, three dice are agitated in a cage shaped like a hourglass. A player may wager upon any of the outcomes 1 through 6. If precisely one die exhibits that value, the player wins at even odds; if two dice show that value, the player wins at 2 to 1 odds; if all three dice show the player's choice, the payoff is 3 to 1. If none of the three dice show the

player's choice, the player loses. Compute the expected value of the player's winnings on a bet of one dollar on "2". Is the game fair? If not, suggest payoff odds that would make the game fair.

14. What is the average number of dots shown by a die tossed once at random? You wish to maximize the value shown by the die. If you are allowed to throw the die a second time, when should you do so? What is the expected value shwon by a die for which one is allowed one rethrowing?

15. James Bond is imprisoned in a cell from which there are three obvious ways to escape: an air-conditioning duct, a sewer pipe and the door (the lock of which doesn't work). The air-conditioning duct leads agent 007 on a two-hour trip whereupon he falls through a trap door onto his head, much to the amusement of his captors. The sewer pipe is similar but takes five hours to traverse (it takes longer to swim then to crawl even for James Bond). Each fall produces temporary amnesia and he is returned to the cell immediately ofter each fall. Assume that he always immediately chooses one of the three exits from the cell with probability 1/3. On the average how long does it take before he notices that the door is unlocked?

16. As new engines are coming off the assembly line in Detroit, they are tested to determine the maximum deliverable horsepower. In a lot of 50 engines, 49 deliver a maximum of 200 horsepower, while one of them doesn't work at all thereby delivering a maximum of 0 horsepower. What is the average maximum horsepower of the

engines in the lot? Is the average a reasonable description of the maximum horsepower of the engines in the lot?

17. A gambler hits upon what seems to be a foolproof system. He begins with a one-dollar bet playing the game of black-or-Red in Roulette, and each time that he loses he doubles the amount bet over the previous bet until he wins *once* at which point he quits. In this way he stands to recoup his losses when he finally does win. He realizes that there is a small chance that he will lose everything he has ($1023), but he considers this probability to be small enough that he can ignore it. The probability of winning on a given trial is 18/36, in which case he wins an amount equal to what he originally bet, otherwise he loses his bet. What is the probability that he eventually wins and what is his net gain when he does? What is the probability that he loses all and how much does he lose? What is his average net gain using this system? Is it really foolproof? Is the risk he is taking a reasonable one?

18. Compute the average length of a Craps game. For the rules of the game see exercise II.20.

In the remaining problems of this dection one will need a hand calculator. In addition we mention the following very useful formula known as *Euler's approximation to the harmonic series*:

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \approx \ell n(n) + 0.57721 \ldots. \quad ,$$

where ℓn denotes the natural logarithm ($\log_e$) and $0.57721 \cdots\cdots$ is known as Euler's constant.

19. A young baseball fan wants to collect a complete set of 262 baseball cards. The baseball cards are available in a completely random fashion, one per package of chewing gum, which she buys twice a day. How long on the average does it take her to get the complete set?

20. A super power has 262 missiles stored in well separated silos. An enemy is considering a sneak attack. However, for the attack to succeed every one of the missiles must be destroyed (the missiles are MIRVed: each has 5 independent warheads). We will consider this problem later, but for now we consider the following simple model. Assume each attacking warhead hits one of the enemy missiles with each enemy missle being equally likely to be the one that is hit. How many warheads on the average will be needed to ensure the destruction of every enemy missile?

21. The analysis in exercise 20 is overoptimistic for several reasons. There is a significant probability that a given warhead will hit more of the silos. Furthermore we want no the average number of warheads required but rather the number of warheads needed to ensure with very high probability (say 99%) that all the enemy missiles have been destroyed. Compute the number of warheads needed if each attacking warhead has probability 0.75 of hitting its target? Even this is optimistic inasmuch as the shock waves form nuclear explosions are such that one cannot expect the various warheads converging on one target or on nearby targets to be independent. However, it gives one an idea of just how foolhardy so-called pre-emptive warfare can be.

22. A molecular beam is firing metal ions toward the face of a crystal. If an ion strikes an unoccupied site on the crystal, it promptly occupies that site, otherwise it bounces away and is lost. Every ion hits the crystal somewhere with each site being equally likely. If there are 10^{16} crystal sites, how many ions must the beam fire at the crystal, on the average, in order to fill every site?

23. Guests arrive at random at a party, and the host seats them as they arrive successively one at a time around a large circular table. Twenty guests arrive, ten single men and ten single women. On the average how many of the twenty adjacent pairs around the table will consist of a man and a woman?

24. The host in exercise 23 invites twenty couples to a cocktail party. As the couples do not know each other, the host decides to mix his guests by assigning each man to one of the women in such a way that every possible arrangement is equally likely. How many couples on the average find themselves assigned to each other? See exercise 58.

25. Return to exercise II.9. How many people on the average call out their birthdays before a match is found, assuming that one is eventually found? How does it compare with the observed value?

26. The Polish mathematician Banach kept two match boxes, one in each pocket. Each box initially contained n matches. Whenever he wanted a match he reached into one of his pockets completely at random. When he found that the box he chose was empty, how many matches were in the other box? how many were there on the average?

27. Compute the average energy of the configuration in exercise II.30.

28. James Bernoulli proposed the following dice game. The player pays one dollar and throws a single die. He then throws a set of n dice, where n is the number shown by the first die. The total number of dots shown by the n dice is then used to determine the payoff.

If the number is less than 12 he loses the bet, if the number equals 12 his dollar is returned, while if the number exceeds 12 he receives two dollars. Find the expected number of dots shown by the n dice. Is the game favorable to the player?

29. Nicolas Bernoulli proposed the following coin-tossing game which has since been called the St. Petersburg paradox. A player pays an entrance fee of E rubles to the casino. A coin is then tossed until it comes up heads. If it requires n tosses to get the first head, the player is paid 2^n rubles, for a net gain of $2^n - E$ rubles. What is the player's expected net gain? [Answer: infinite net gain no matter how large E is] The paradox arises from the fact that one is placing no limit on the resources of the casino. If the casino possesses a total of $P = 2^N$ rubles, compute the net expected gain of the player. For the game to be fair what should E be? [Answer: $N+1$ rubles]. For example, if the casino has resources of 33.55 million rubles, what entrance fee would be fair?

30. What is the expected duration of the St. Petersburg game for the casino mentioned at the end of exercise 29?

31. What is the probability that in n tosses of a fair coin two heads never occur in a row, i.e. no run of 2 or more heads ever occurs?

32. The generalization of Chevalier de Méré's first problem (exercise II. 27) is called the problem of points. The problem concerns a game between two players that was interrupted before its conclusion. Suppose that N points are required to win the game, that player A has $N-a$ points and that player B has $N-b$ points. In a given trial A wins with probability p and B with probability $q = 1 - p$. How should the stakes be divided? The problem was first solved by Montmort. Can you solve it also?

33. Generalize exercise 14 to produce a kind of analog, for dice throwing, of draw poker. The player throws five dice. He then has the option to choose a subset of the dice for rethrowing. This subset can be empty but cannot consist of all the dice. The process is then repeated for the rethrown dice, continuing until no more dice may be rethrown. The object is to maximize the total number of dots showing on the dice. Devise a strategy and calculate the expected outcome for this strategy. The optimal strategy will produce an expected outcome of about $24 \frac{4}{9}$.

Continuous Random Variables

34. A boy makes a date with his girl friend. They are to meet at some time between 6 PM and 7 PM, but since both are absent-minded they forget which time they had agreed upon. As a result

each arrives at a random moment between 6 and 7. Each waits for 10 minutes and if the other fails to appear, he or she promptly leaves in a blue funk. What is the probability that true love prevails (at least this one evening)?

35. When five points are chosen uniformly at random from the interval [1,2], what is the distribution of the natural logarithm of the smallest point?

36. A gangster stands 10 m from an infinitely long straight wall. The gangster fires a gun horizontally in a completely random <u>direction</u> toward the wall. Compute the distribution of the point on the wall where the bullet hits. Do the same for the distance from the bullet to the point of the wall closest to the gangster.

37. A <u>median</u> of a random variable X is any number μ such that

$$P(X \leq \mu) \geq \frac{1}{2} \quad \text{and} \quad P(X \geq \mu) \geq \frac{1}{2} \;.$$

Prove that the median of any random variable exists. Does it have to be unique? Compute it for the gangster distribution in exercise 36 above.

38*. After grading an examination, a teacher arranges the papers in order by grade. The <u>sample median</u> is the middle grade if there are an odd number of papers and is the average of the two middle grades otherwise. Give a definition of the sample median in the Uniform process and compute its distribution.

39. How far apart are the largest and the smallest points in the Uniform process of sampling n points from $[0,a]$? We call this the spread. Compare the spread with the second largest order statistic, $X_{(n-1)}$.

40. Show that any function satisfying the four properties of a distribution function is in fact the distribution of some random variable on some sample space.

41.* One can also develop a theory of discrete order statistics. For this the interval $[0,a]$ is replaced by the set of integers $\{1, 2, \cdots, A\}$, each of which is equally likely to be chosen, and a given integer may be chosen more than once. The formulas one gets are quite complicated. It should be clear, however, that when A is large compared to n, the number of points chosen, we may approximate the discrete order statistics with the continuous ones. The principle that the gaps are equidistributed holds both for the discrete and for the continuous cases. Compute the distribution of the first order statistic. Note that this is an integer random variable.

42.* During World War II, the Allies estimated the number of tanks that had been produced by German industry by collecting the serial numbers of abandoned tanks. There are actually two questions one can ask here. One can ask for the most likely number of tanks that have been produced, or one can ask for the most reasonable rough estimate of the number. The former question would be most appropriate if we placed a very high value on getting the exact number, nearby numbers being useless. The latter question is clearly more appropriate in the context of this problem.

To answer these questions we must rephrase them in the language of probability. Assume n serial numbers have been collected, the largest of which is $X_{(n)}$. The first question should read: what is the number A of tanks such that when n numbers are chosen uniformly from $\{1, \cdots, A\}$ the probability that we get the actually observed values is as large as possible. The answer is $X_{(n)}$ itself. Prove this. We call this the _maximum likelihood estimator_ of A . See exercise II.23 for another example of such an estimator. For the second question we want an estimate of A such that if one makes many estimates of the same number A by this method we will on the average be close to the correct value. We will consider this question later in exercise 52.

43.* A biologist is studying organelles in a cell. The organelles in question are spheres of equal but unknown radius r within a given type of cell, and they are distributed randomly throughout the cell. The biologist estimates r by observing a cross-section of the cell and measuring the radii of the visible granules. Suppose that n granules are observed and that the largest observed radius is $R_{(n)}$. Fine the maximum likelihood estimator of r . The measurements $R_1, \cdots, R_n$ of the n radii will not be uniformly distributed. However the random variables $\sqrt{r^2 - R_1^2}, \ldots, \sqrt{r^2 - R_n^2}$ will be uniformly distributed.

44.* Compute directly, without first finding densities, the joint distribution function of the two order statistics $X_{(j)}$ and $X_{(k)}$ in the Uniform process.

45.* Suppose that $X_1, X_2, \ldots, X_n$ are independent uniformly distributed random variables on the intervals $[0,a_1], \ldots, [0,a_n]$ respectively. Compute the densities and the joint densities of the order statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$.

46.* In exercise 45 above, compute the distributions of the gaps.

Expectations of Continuous Random Variables

47. Compute the average value of the natural logarithm of the smallest point among five chosen uniformly from [1,2]. Is this the same as the natural logarithm of the average value of the smallest point? Explain this. See exercise 35.

48. Compute the average of the median of the set of order statistics. See exercise 38.

49. Compute the average values of the random variables in exercise 36 (the gangster distributions).

50. An enzyme randomly breaks each of 24 identical (and very long) DNA molecules into two pieces. How long is the shortest piece produced? What is the average length of the shortest piece?

51. Shuffle a deck of cards and turn up cards one at a time until the first spade appears. How many cards including the spade do we expose on the average? (Answer: $\frac{53}{14} \approx 3.786$) More generally if we are looking for one of n cards in the deck, how many cards must we expose on the average until we find one of them? (Answer: $\frac{53}{n+1}$)

52.* Return to exercise 42. To answer the second question we require a random variable with the property that its expectation is A. The maximum likelihood estimator will not do because $E(X_{(n)}) \neq A$. What should one use? $\left[\text{Answer: } \frac{n+1}{n} X_{(n)}\right]$

Consider next the corresponding question for exercise 43. The situation is now more complicated because the observed radii are not uniformly distributed. Find a random variable R for this problem such that $E(R) = r$. $\left[\text{Answer: } \frac{n}{\sqrt{n^2 - 1}} R_{(n)}\right]$.

53.* A long DNA molecule is broken into N pieces. Find the average length of the i^{th} longest piece produced, $1 \leq i \leq N$. Use probabilistic reasoning as follows. Let $n = N - 1$ so that our model is the Uniform process of sampling n points from $[0,a]$, where a is the length of the DNA molecule. The problem is to compute the expectations of $L_{(1)}, L_{(2)}, \cdots, L_{(n+1)}$, i.e. of the order statistics of the gaps. Here is how to compute $L_{(1)}$. First find $P(L_{(1)} > t)$. Now $(L_{(1)} > t)$ is the event $(L_1 > t) \cap (L_2 > t) \cap \cdots \cap (L_{n+1} > t)$. When this event occurs we can remove a subsegment of length t from each of the gaps. The resulting process in the Uniform process of sampling n points from $[0, a - (n+1)t]$. Geometrically the event $(L_{(1)} > t)$ is an n-dimensional cube having side of length $a - (n+1)t$. Thus $P(L_{(1)} > t) = \left(\frac{a - (n+1)t}{a}\right)^n$. Therefore $\text{dens}(L_{(1)} = t) = n(n+1) \frac{(a - (n+1)t)^{n-1}}{a^n}$. It is now easy to calculate $E(L_{(1)})$ $\left[\text{Answer: } \frac{a}{(n+1)^2}\right]$. Similarly to compute $L_{(2)}$ we simply note that when we remove a subsegment of length

$L_{(1)}$ from each gap (which we can do since $L_{(1)}$ is the smallest gap), what remains is the Uniform process of sampling $n-1$ points from $[0, a-(n+1)L_{(1)}]$. Moreover $L_{(2)}$ is the sum of $L_{(1)}$ and the length of the smallest gap in this smaller process. This gives $E(L_{(2)})$. $\left[\text{Answer: } E(L_{(2)}) = E(L_{(1)}) + E\left\{\frac{a-(n+1)L_{(1)}}{n^2}\right\}\right.$

$$= \frac{a}{(n+1)^2} + \frac{a}{n^2} - \frac{(n+1)a}{n^2(n+1)^2} = \frac{a}{(n+1)^2} + \frac{a}{n(n+1)} = \left.\frac{a}{n+1}\left(\frac{1}{n+1} + \frac{1}{n}\right)\right].$$

Continuing by induction we get $E(L_{(i)})$ for all i. $\left[\text{Answer: } E(L_{(i)}) = \frac{a}{n+1}\left(\frac{1}{n+1} + \cdots + \frac{1}{n-i+2}\right)\right]$. Although the above reasoning is not, strictly speaking, rigorous, we will show how to make it completely rigorous in Chapter V. See exercise V.24.

54.* A biologist allows an enzyme to break a DNA molecule into 10 pieces. The original molecule was 10,000 base pairs long. Upon examining the pieces, the biologist finds that the smallest is only 10 base pairs long! How probable is it that the smallest of 10 pieces could be this short or shorter? Use the results of exercise 53. Does the biologist have a case for believing that the enzyme attacked the DNA molecule in a non-random fashion? [Answers: 8.6%; the event is not at all surprising]

The Inclusion-Exclusion Principle

55. A gambler is playing a sequence of games. For each trial he can choose to bet either on heads or on tails of a toss of a fair coin. If he bets on heads he gains or loses $1 depending on whether the coin shows heads or tails respectively. Similarly if he bets on tails he gains or loses $2 if the coin shows tails or heads respectively. In each trial the gambler chooses one or the other bet at random betting on heads with probability p.

Let A be the event that he bets on heads and let B be the event that the coin shows heads. Write his net gain in one trial using indicators. [Answer: $I_A I_B - I_A I_{\bar{B}} + 2I_{\bar{A}} I_{\bar{B}} - 2I_{\bar{A}} I_B$]

56. The President of the U.S. holds frequent news conferences. The journalists who attend these conferences are usually the same group, more or less. Let us suppose that in the first two years of his term the President answers 400 questions put to him by the 100 regular journalists. During this time 4 of the journalists have never been accorded recognition. These four get together and complain that they are being discriminated against, arguing that the probability that none of them was ever recognized is only 8×10^{-8}. On the other hand, the President's Press Secretary argues that the probability for four or more of the journalists to be ignored is really about 11%, which is not significant evidence for discrimination. Who is right? Formulate the two models being used and calculate the required probabilities. Refer to exercise II.17 for one model. For the other use the finite uniform process of sampling with replacement 400 journalists from among 100. The latter calculation requires a small computer. In Chapter VI we will develop techniques for approximating the answer with much less effort. See exercise VI.31.

57. In the game of Treize, popular in seventeenth century France, 13 balls labelled from 1 through 13 were placed in an urn and drawn out one at a time at random without replacement. The players bet on the waiting time until either the n^{th} ball drawn was labelled n or else the urn was emptied. Compute the distribution of this waiting time. Generalize to N balls.

58. An obvious modification of the game of Treize would be to allow players to bet on the number of times that the n^{th} ball drawn was labelled n. Compute the distribution of this number. Generalize to N balls. What is the average number of matches?

59. A sociologist claims that he can determine a person's profession by a single glance. A psychologist decides to test his claim. She makes a list of 13 professions and chooses photographs (all in a standard pose) of 13 individuals one in each profession. She then asks the sociologist to match the photos with professions. The sociologist identifies only 5 correctly. What do you think of his claim? Note that this exercise is closely related to exercise 58.

60.* Prove the inclusion-exclusion principle for max and min.

$$\text{Hint:} \quad \int_c^\infty I_{(-\infty,x)}\, dy = x - c \,, \quad \text{provided} \quad c < x \,.$$

61.* Return to the molecular beam in exercise 22. Assume the beam fires 10^{14} ions per second. Compute the distribution of the waiting time until the crystal is totally covered. Give a formula. Don't try to evaluate it. In exercise VI.32 we will show how to compute an accurate approximation for the value of this expression.

Do the same computation as above for the baseball fan problem (exercise 19) and for the pre-emptive nuclear attack problem (exercise 20).

62.* In a physical configuration there are b bosons, each in one of N states. Compute the distribution of the number of filled states (i.e. the number of states having one or more bosons).

63.* The standard card deck used by ESP experimenters is called the Zener deck. It has twenty-five cards, five each of five kinds. A typical test consists of the experimenter in one room and the subject in another. The experimenter shuffles the deck thoroughly and then turns the cards over one at a time at a fixed rate. Simultaneously, the subject is trying to perceive the sequence of cards. In order to test whether the subject's perceived sequence could have been simply a random guess, we must calculate the distribution of the number of matches occurring in a random permutation of the deck relative to some standard ordering of the deck.

64.* In an ancient kingdom the new monarch was required to choose his queen by the following custom. One hundred prospective candidates are chosen from the kingdom and once a day for one hundred days one of the candidates chosen at random was presented to the monarch. The monarch had the right to accept or reject each candidate on the day of her presentation. When a given candidate is rejected she immediately gets married, and so the monarch cannot change his mind. Assume that the preferences of the monarch can be expressed in a linear order (from the best to the worst) and that the monarch wants the best candidate, second best won't do. What strategy should he employ? What is the probability that he succeeds in his quest for perfection?

Suppose that instead of simply ranking the candidates, the monarch rates each of them on a 0 to 10 scale, i.e. using some sort of objective criteria, he computes a real number between 0 and 10 for each. Assume that the ratings are uniformly distributed on [0,10]. Again the monarch wants the best candidate among the 100. What strategy should he employ now? What is the probability of success?

Chapter IV Statistics and the Normal Distribution

The normal distribution arises whenever we make a succession of imperfect measurements of a quantity that is supposed to have a definite value. If all the students in a class take the same test, we may think of their grades as being imperfect measurements of the average capability of the class. In general, when we make a number of independent measurements, the average is intuitively going to be an approximation of the quantity we are trying to find.

On the other hand, the various measurements will tend to be more or less spread out on both sides of the average value. We need a measure for how far individual measurements are spread out around the quantity being measured. This will tell us, for example, how many measurements must be made in order to determine the quantity to a certain accuracy. It will also make it possible to formulate statistical tests to determine whether or not the data in an experiment fit the model we have proposed for the experiment.

1. Variance

The variance of a random variable X is a measure of the spread of X away from its mean.

Definition. Let X be a random variable whose mean is $E(X) = m$. The variance of X is $\mathrm{Var}(X) = E((X-m)^2)$, if this expectation converges. The square root of the variance is called the standard deviation of X and is written $\sigma(X) = \sqrt{\mathrm{Var}(X)}$. We sometimes write $\sigma^2(X)$ for Var(X).

If X is an integer random variable having probability distribution $p_n = P(X=n)$, then

$$\mathrm{Var}(X) = \sum_n (n-m)^2 p_n.$$

If X is a continuous random variable whose density is $f(x) = \mathrm{dens}(X=x)$, then

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-m)^2 f(x)\,dx.$$

In the continuous case we can imagine that f(x) is the density at x of a thin rod. This rod has total mass 1 and balances at the mean m. The moment of inertia of this rod about its balance point is precisely the variance. If we rotate the rod about m it would have the same angular momentum if all the mass were concentrated at a distance $\sigma(X)$ from the point of rotation.

It is possible for a random variable not to have a mean. It is also possible for a random variable to have a mean but not to have a finite variance. We shall see examples in the exercises. However in a great many physical processes it is reasonable to assume that the random variables involved do have a finite variance (and hence also a mean). For example, on an exam if the possible scores range from 0 to 100, the measurement of someone's exam score is necessarily going to have a finite variance.

A useful formula for the variance is the following:

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

This is an easy formula to verify. The crucial step is that the expectation is additive, even when the random variables involved are not independent.

$$
\begin{aligned}
\mathrm{Var}(X) &= E(X-m)^2) \\
&= E(X^2-2mX+m^2) \\
&= E(X^2)-2mE(X)+m^2 \\
&= E(X^2)-2m^2+m^2 \\
&= E(X^2)-m^2.
\end{aligned}
$$

As we just remarked, the expectation is additive. In general it is not multiplicative; that is, $E(XY)$ need not be $E(X)E(Y)$. The variance is a measure of the extent to which the expectation is not multiplicative when $X = Y$, for in this case we have $\mathrm{Var}(X)=E(X\cdot X)-E(X)E(X)$. The *covariance* of X and Y in general is the difference

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

We will not be using covariances except in a few optional exercises. Covariances are often used as a measure of the independence of random variables because of the following important fact:

Fact. If X and Y are independent random variables, then E(XY) = E(X)E(Y), or equivalently Cov(X,Y) = 0.

Proof. We will only consider the case of integer random variables. The proof for the continuous case requires annoying technicalities that obscure the basic idea. We leave these as an exercise.

We compute the distribution of the product XY in terms of the distributions of X and of Y using the law of alternatives and the fact that X and Y are independent.

$$\begin{aligned} P(XY=n) &= \sum_k P(XY=n \mid X=k)P(X=k) \\ &= \sum_k P(Y=n/k \mid X=k)P(X=k) \\ &= \sum_k P(Y=n/k)P(X=k). \end{aligned}$$

Therefore the expectation of XY is

$$\begin{aligned} E(XY) &= \sum_n nP(XY=n) \\ &= \sum_n n \sum_k P(Y=n/k)P(X=k) \\ &= \sum_n \sum_k nP(Y=n/k)P(X=k). \end{aligned}$$

Finally change variables to j and k where j = n/k. Then

$$
\begin{aligned}
E(XY) &= \sum_j \sum_k jkP(Y=j)\ P(X=k) \\
&= \sum_j jP(Y=j) \sum_k k\ P(X=k) \\
&= E(X)E(Y). \quad \text{This completes the proof.}
\end{aligned}
$$

We add that it is possible for non-independent random variables X and Y to satisfy $E(XY)=E(X)E(Y)$. As a result the covariance is not a true measure of independence.

Now whereas the expectation is additive whether the random variables are independent or not, the variance need not be additive in general. The most important consequence of the above fact is that for independent random variables X and Y, the variance is additive.

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

In terms of the standard deviations, $\sigma(X+Y) = \sqrt{\sigma^2(X)+\sigma^2(Y)}$; the standard deviations of independent random variables act like the components of a vector whose length is $\sigma(X+Y)$.

There are two more properties of the variance that are important for us. Both are quite obvious:

$$\text{Var }(cX) = c^2 \text{ Var}(X) \qquad\qquad \text{Var}(X+c) = \text{Var}(X).$$

The first expresses the fact that the variance is a quadratic concept. The second is called <u>shift invariance</u>. It should be obvious that merely shifting the value of a random variable X by a constant only changes the mean and not the spread of the measurement about the mean.

(0) $\text{Var}(X) = E(X^2)-E(X)^2 \qquad \sigma(X) = \sqrt{\text{Var}(X)}$

(1) If X and Y are independent random variables having finite variance, then:

$$\text{Var}(X+Y) = \text{Var}(X)+\text{Var}(Y) \qquad \sigma(X+Y) = \sqrt{\sigma^2(X)+\sigma^2(Y)}$$

(2) $\text{Var}(cX) = c^2\text{Var}(X) \qquad \sigma(cX) = c\sigma(X)$

(3) $\text{Var}(X+c) = \text{Var}(X) \qquad \sigma(X+c) = \sigma(X)$

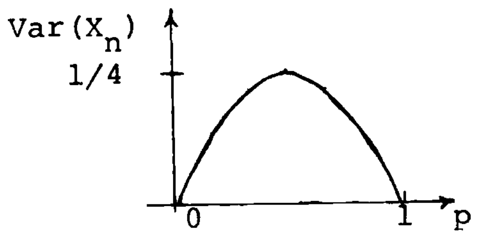
Basic Properties of Variance and Standard Deviation

We now compute the variances of some of the random variables we have encountered so far in the Bernoulli, Uniform and Poisson processes.

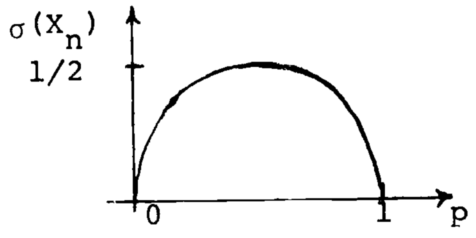
Bernoulli Process

Consider a single toss of a biased coin. This is described by any random variable X_n of the Bernoulli process. Recall that $X_n = \begin{cases} 0 & \text{if } n^{th} \text{ toss is tails} \\ 1 & \text{if } n^{th} \text{ toss is heads} \end{cases}$

Since $1^2 = 1$ and $0^2 = 0$, X_n^2 is the same as X_n. Therefore the variance of any X_n is $\mathrm{Var}(X_n) = E(X_n^2)-E(X_n)^2 =$ $E(X_n)-E(X_n)^2 = p-p^2 = pq$.



The variance of a toss of a coin whose bias is p.



The standard deviation of a toss of a coin whose bias is p.

Notice that the largest variance corresponds to a fair coin (p=1/2). We intuitively think of a fair coin as having the most "spread out" distribution of all biased coins; while the more biased a coin is, the more its distribution is "concentrated" about its mean.

Next consider the number of successes S_n in n tosses of **a biased coin.** Since $S_n = X_1+X_2+...+X_n$ is the sum of n independent random variables all of whose variances are the same, $\mathrm{Var}(S_n) = n\mathrm{Var}(X_1) = npq$. If we tried to compute $\mathrm{Var}(S_n)$ directly from the definition, we get

$$\mathrm{Var}(S_n) = E((S_n-np)^2) = \sum_{k=0}^{n} (k-np)^2 \binom{n}{k} p^k q^{n-k} .$$

That this is npq is far from obvious.

We leave the computation of the variances of the gaps T_k and the waiting times W_k as exercises.

$$\mathrm{Var}(T_k) = \frac{q}{p^2} \qquad\qquad \mathrm{Var}(W_k) = \frac{kq}{p^2}$$

Uniform Process

Consider a point X dropped at random uniformly on [0,a]. Clearly the average value of the point is a/2, the midpoint of [0,a]. The variance is

$$\begin{aligned}\mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= \int_0^a x^2 \,\mathrm{dens}(X{=}x)\,dx - \left(\frac{a}{2}\right)^2 \\ &= \int_0^a x^2 \frac{1}{a}\,dx - \left(\frac{a}{2}\right)^2 \\ &= \frac{1}{a}\left[\frac{x^3}{3}\right]_0^a - \frac{a^2}{4} \\ &= \frac{1}{a}\,\frac{a^3}{3} - \frac{a^2}{4} \\ &= \frac{a^2}{12}\,.\end{aligned}$$

So the standard deviation is $\frac{a}{\sqrt{12}} = \frac{a}{2\sqrt{3}} \approx 0.2887a$. We can think of this in the following way. Given a uniform bar of length a, its midpoint is the center of mass. If the bar were set spinning around its center of mass, the angular momentum would be the same if the mass were all at a distance $\frac{a}{2\sqrt{3}}$ from the center of rotation.

We leave it as an exercise to compute the variances of the gaps and the order statistics of the Uniform process. Notice that we cannot use the fact that $X_{(k)} = L_1+L_2+\ldots+L_k$ because the gaps are <u>not</u> independent.

$$\mathrm{Var}(L_i) = \frac{a^2 n}{(n+1)^2(n+2)}$$

$$\mathrm{Var}(X_{(k)}) = \frac{a^2 k(n-k+1)}{(n+1)^2(n+2)}$$

We summarize the above computations in this table.

Distribution	Model(s)	Expectation	Variance
Bernoulli	X_i in the Bernoulli process	p	pq
Binomial	S_n (X_i when $n=1$) in the Bernoulli process	np	npq
Geometric	W_1 or any T_k in the Bernoulli process	$1/p$	q/p^2
Negative Binomial	W_k in the Bernoulli process	k/p	$^{kq}/_{p^2}$
Uniform	X_i in the Uniform process	$a/2$	$a^2/12$
Dirichlet	$X_{(k)}$ (L_i when $k=1$) in the Uniform process	$ka/(n+1)$	$\frac{a^2 k(n-k+1)}{(n+1)^2(n+2)}$

Table of Means and Variances

Standardization

If we shift a random variable X by a constant, replacing X by X+c or if we make a scale change, multiplying X by a nonzero constant, we have not altered X in a significant way. We have only reinterpreted a measurement of X by a linear change of variables. The idea of standardization is to choose a single "standard" random variable among all those related to one another by a linear change of variables. Then in order to determine if two random variables are "essentially" the same we should compare their standardized versions.

Definition. A random variable X is *standard* or *standardized* if E(X) = 0 and Var(X) = 1. If X has finite variance, then

$$\frac{X-m}{\sigma},$$

where m = E(X) and $\sigma = \sigma(X)$, is standard. We call $(X-m)/\sigma$ the *standardization* of X. A physicist would say that $\frac{X-m}{\sigma}$ expresses X in "dimensionless units."

The covariance of the standardizations of two random variables X and Y is called the *coefficient of correlation* and is written $\rho(X,Y)$. It is easy to prove that $|\rho(X,Y)| \leq 1$ and that $\rho(X,Y) = \mathrm{Cov}(X,Y)/(\sigma(X)\sigma(Y))$, and we leave these as exercises. Because of the importance of standardization, we will prove that the standardization of a random variable is really standard.

Fact. If X has finite variance, then $(X-m)/\sigma$ is standard.

Proof

$$E((X-m)/\sigma) = \frac{E(X)-m}{\sigma} = \frac{m-m}{\sigma} = 0$$

$$\begin{aligned}\mathrm{Var}((X-m)/\sigma) &= \frac{1}{\sigma^2}\,\mathrm{Var}(X-m) && \text{(Basic fact 2)}\\ &= \frac{1}{\sigma^2}\,\mathrm{Var}(X) && \text{(Basic fact 3)}\\ &= 1 && (\text{since } \sigma^2 = \mathrm{Var}(X))\end{aligned}$$

We call $\sigma(X)$ the *standard* deviation because of its appearance in the standardization. We think of $\sigma(X)$ as being the natural unit for measuring how far a given observation of X deviates from the mean. The importance of standardization will gradually emerge in the next few sections.

2. The Bell-Shaped Curve

In this section we introduce one of the most important distributions in probability: the normal distribution. The traditional explanation for the importance of the normal distribution relies on the Central Limit Theorem, which we will discuss in the next section. However we feel that the explanation, using entropy and information, given in chapter VII is better because it provides a context which explains the ubiquity not only of the normal distribution but of several other important distributions as well.

Definition. A continuous random variable X is said to have the *normal* or *Gaussian distribution with mean* m *and variance* σ^2 if

$$\text{dens}(X=x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}$$

For brevity we will write simply "X is $N(m,\sigma^2)$". Some authors write $N(m,\sigma)$ instead of $N(m,\sigma^2)$; one should beware.

Unlike most distributions, the formula for the normal density comes with the mean and standard deviation already specified. We should, however, verify that m and σ^2 really are the mean and variance. In fact it isn't obvious that this formula actually defines the density of a continuous random variable. To verify these facts we use the following basic formula, which everyone ought to have seen in a calculus course at some time:

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} .$$

To prove this let $A = \int_{-\infty}^{\infty} e^{-x^2} dx$. Then, since x is a dummy variable, $A = \int_{-\infty}^{\infty} e^{-y^2} dy$ also. Therefore $A^2 = \int_{-\infty}^{\infty} e^{-x^2} dx \int_{-\infty}^{\infty} e^{-y^2} dy$. Now we switch to polar coordinates and integrate:

$$\begin{aligned} A^2 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dy \\ &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2} rdrd\theta \\ &= \int_0^{2\pi} \left[-\frac{1}{2} e^{-r^2} \right]_0^{\infty} d\theta \\ &= \int_0^{2\pi} \frac{1}{2} d\theta = \pi . \end{aligned}$$

Hence $A = \sqrt{\pi}$.

We use this formula first to show that the normal density really defines a density. We first change variables to

$y = (\frac{x-m}{\sigma\sqrt{2}})$ so that $\sigma\sqrt{2}\,dy = dx$. Then

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}\,dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-y^2}\,\sigma\sqrt{2}\,dy$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2}\,dy$$

$$= 1.$$

Next we compute $E(X)$ when X is $N(m,\sigma^2)$. Again we use the change of coordinates $y = \frac{x-m}{\sigma\sqrt{2}}$. Note that $x = \sigma\sqrt{2}\,y+m$.

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}\,dx$$

$$= \int_{-\infty}^{\infty} (\sigma\sqrt{2}y+m)\cdot\frac{1}{\sigma\sqrt{2\pi}} e^{-y^2}\,\sigma\sqrt{2}\,dy$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} (\sigma\sqrt{2}\,y+m)\,e^{-y^2}dy$$

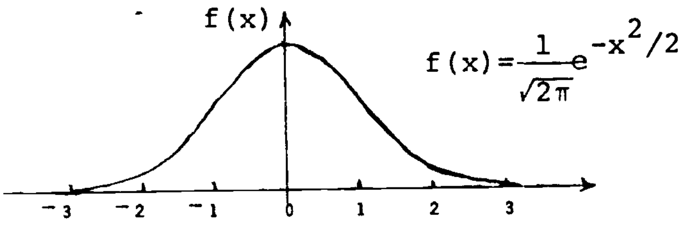
$$= \frac{\sigma\sqrt{2}}{\sqrt{\pi}} \int_{-\infty}^{\infty} y\,e^{-y^2}\,dy + \frac{m}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2}\,dy$$

$$= \frac{1}{\sqrt{\pi}} \left[-\frac{1}{2} e^{-y^2}\right]_{-\infty}^{\infty} + \frac{m}{\sqrt{\pi}} \sqrt{\pi}$$

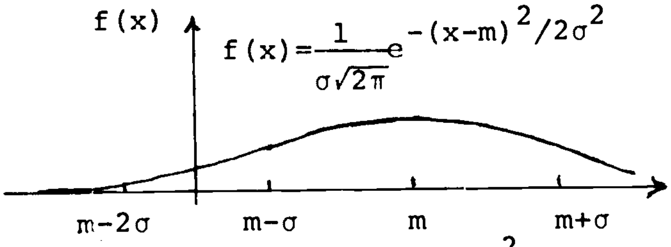
$$= 0 + m.$$

Finally we leave it as an exercise to show that Var(X) = σ^2. It can be done using integration by parts.

Although these computations look messy, we are inescapably forced to consider this density function because this is the distribution corresponding to the concept of total randomness or complete randomness. In Chapter VII we will make this concept more precise. So it is important that one have an intuitive idea of what it means for a random variable to be normal. We suggest that the following properties of the normal distribution be memorized, and one should familiarize oneself with the use of the tables giving values of the normal distribution function.

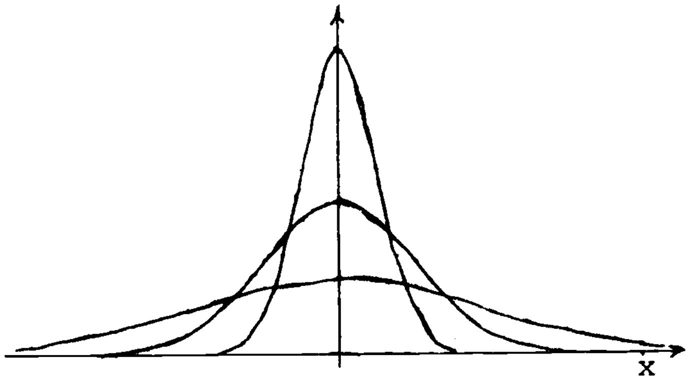


The standard normal density N(0,1)



The normal density $N(m,\sigma^2)$

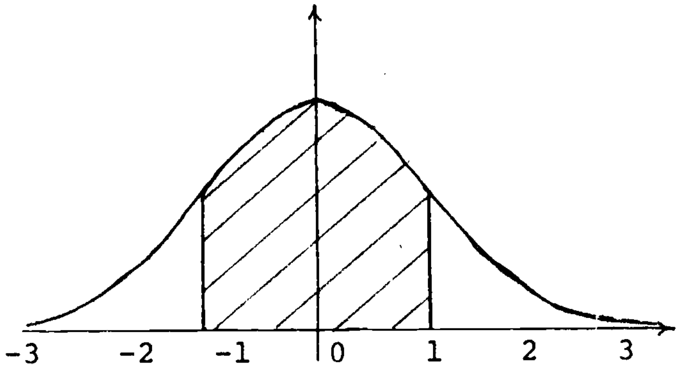
The normal density function is symmetric about the mean m and the maximum value is taken at the mean. Beyond 3.5σ units from m, the value of the normal density is essentially zero. The natural unit for measuring deviations from the mean is the standard deviation. When x is the deviation from the mean measured in this natural unit, then we get the standard normal density. Most tables for the normal distribution are tables of the standard normal density.



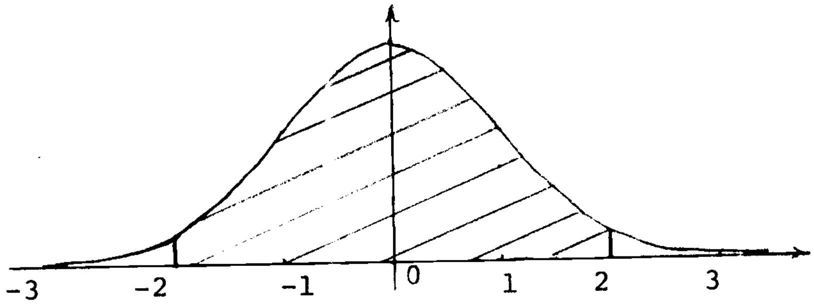
The curve becomes steeper and higher at the mean as σ gets smaller.

Various normal densities with mean 0

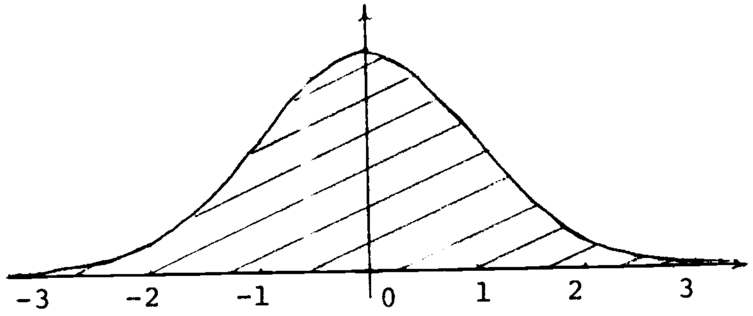
In all of the following X is standard normal, N(0,1).



The area within one standard deviation of the mean is 68.27% of the total area, i.e. $P(-1 \leq X \leq 1) = .6827$.



The area within two standard deviations of the mean is 95.45%, of the total area, i.e. $P(-2\leq X\leq 2) = .9545$.



The area within three standard deviations of the mean is 99.73% of the total area, i.e. $P(-3\leq X\leq 3) = .9973$.

In addition one should memorize the following two cases:

The area within 1.96 standard deviations of the mean is 0.95

The area within 2.58 standard deviations of the mean is 0.99

These will be important when we compute significance levels.

Occasionally one will see tables of the <u>error function</u>, erf(t). This function is closely related to the normal distribution although it is not the same:

$$\operatorname{erf}(t) = P(|Y|\leq t) = \frac{1}{\sqrt{\pi}}\int_{-t}^{t} e^{-y^2}\,dy \;,$$

where Y has distribution N(0,1/2). If X has the standard normal distribution, then

$$P(-x\leq X\leq x) = \mathrm{erf}(\frac{x}{\sqrt{2}})$$

and

$$P(X\leq x) = 1/2 + \mathrm{erf}(\frac{x}{\sqrt{2}})/2 \ .$$

3. The Central Limit Theorem

The traditional explanation for the importance of the normal distribution relies on the Central Limit Theorem. Briefly, this theorem states that the average of n independent equidistributed random variables tends to the normal distribution no matter how the individual random variables are distributed. The explanation for the ubiquity of the normal distribution then goes as follows. Suppose that X is the random variable representing the measurement of a definite quantity but which is subject to chance errors. The various possible imperfections (minute air currents, stray magnetic fields, etc.) are supposed to act like independent equidistributed random variables whose sum is the total error of the measurement X. Unfortunately this explanation fails to be very convincing because there is no reason to suppose that the various contributions to the total error are either independent or equidistributed. We will have to wait until Chapter VII to find a more fundamental reason for the appearance of the normal distribution. The explanation given there uses the concepts of entropy and information.

Intuitively, the sum of independent, equidistributed random variables is progressively more disordered as we add more and more of them. As a result the standardization of the sum necessarily approaches having a normal distribution as $n\to\infty$. This tendency to become disordered is exhibited even when the random variables are not quite independent and equidistributed. It is this tendency that accounts for the ubiquity of the normal distribution.

Nevertheless the Central Limit Theorem is of importance in probability and statistics, particularly in the theory of hypothesis testing which we will be discussing in the next section. Moreover the proof of the Central Limit Theorem is more difficult than our intuitive justification would lead us to believe. We will now give a precise statement of this theorem. The proof is sketched in section 6.

Suppose that $X_1, X_2, \ldots, X_n, \ldots$ are independent equidistributed random variables whose common mean and variance are $m = E(X_i)$ and $\sigma^2 = \text{Var}\ (X_i)$. Let S_n be the sum $X_1 + \ldots + X_n$. Then the mean of S_n is $E(S_n) = E(X_1) + E(X_2) + \ldots + E(X_n) = nm$, and its variance is $\text{Var}\ (S_n) = \text{Var}\ (X_1) + \text{Var}\ (X_2) + \ldots + \text{Var}(X_n) = n\sigma^2$, since the X_i's are independent and equidistributed. Therefore the standard deviation of S_n is $\sigma(S_n) = \sqrt{\text{Var}(S_n)} = \sqrt{n}\sigma$. Hence the standardization of S_n is

$$Y_n = \frac{S_n - nm}{\sigma\sqrt{n}} = \frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}}$$

The Standardization of a Sum of Independent, Equidistributed Random Variables whose common mean and variance are m and σ^2 respectively.

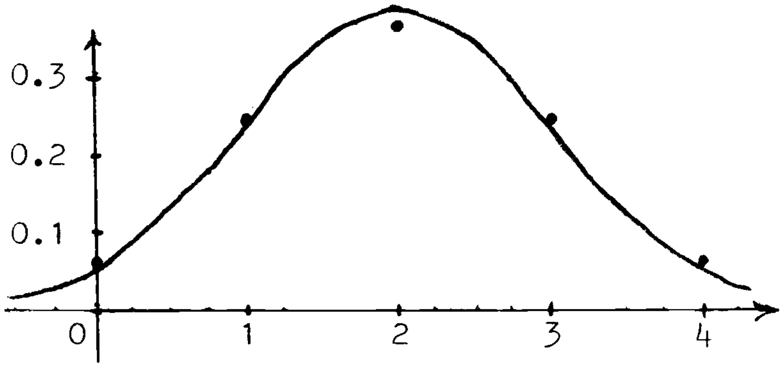
This is an important formula to remember. The Central Limit Theorem then says that Y_n tends toward the normal distribution as $n\to\infty$.

Central Limit Theorem. If $X_1, X_2, \ldots, X_n, \ldots$ are independent equidistributed random variables with mean m and variance σ^2, then

$$P(Y_n \le t) = P\left(\frac{X_1+X_2+\ldots+X_n-nm}{\sigma\sqrt{n}} \le t\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\,dx$$

as $n \to \infty$.

For example, in the Bernoulli process the random variable S_n is the sum of independent equidistributed random variables X_i whose common mean is $m = p$ and whose common variance is $\sigma^2 = pq$. Then $Y_n = \dfrac{S_n - np}{\sqrt{npq}}$ tends toward the standard normal distribution. That is, S_n is approximately distributed according to $N(np,npq)$. This approximation is surprisingly accurate even for small values of n.



The distribution of S_4 in the Bernoulli process using a fair coin. Superimposed is the normal distribution $N(2,1)$

For example, we know that $P(1 \leq S_4 \leq 3) = 1 - \frac{2}{16} = 0.875$. On the other hand, when X is $N(2,1)$, $P(0.5 \leq X \leq 3.5) \simeq 0.8664$ (since 0.5 and 3.5 are each 1.5σ from the mean 2). You can see that the fit is quite close.

The two most common manifestations of the Central Limit Theorem are the following:

(1) As $n\to\infty$, the sum S_n "tends" to the distribution $N(nm, n\sigma^2)$

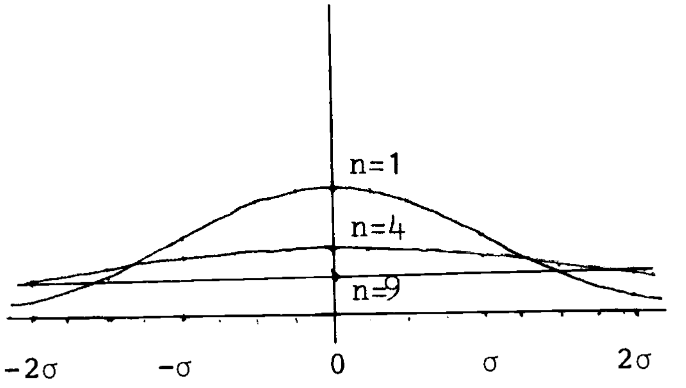
(2) As $n\to\infty$, the sample average $\overline{m}=S_n/n$ "tends" to the distribution $N(m, \sigma^2/n)$.

The expectation of the <u>sample</u> <u>mean</u> is the mean as we already noticed:

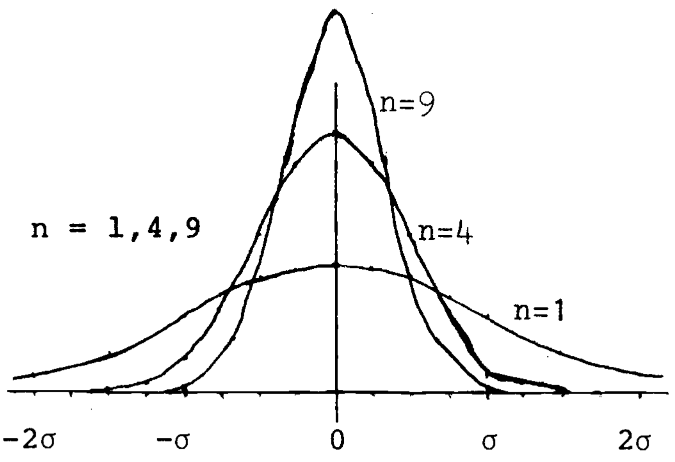
$$E(\overline{m}) = E\left(\frac{S_n}{n}\right) = \frac{nm}{n} = m,$$

hence the sample mean is an approximation to the true mean m. The spread of the sample mean depends on the variance $\mathrm{Var}(\overline{m}) = \mathrm{Var}(S_n/n) = \mathrm{Var}(S_n)/n^2 = n\sigma^2/n^2 = \frac{\sigma^2}{n}$.

Intuitively it is clear that as $n\to\infty$, the sample average will be a better and better approximation to the mean m. The Central Limit Theorem tells us precisely how good an approximation it is. In the following drawings we assume $m = 0$.



As $n\to\infty$, the distribution of S_n tends to broaden. The spread of S_n is proportional to $\sqrt{n}$.



As n→∞, the distribution of $\overline{m}$ tends to become steeper. The spread of $\overline{m}$ is proportional to $1/\sqrt{n}$.

We remark that the independence of the random variables is essential in the Central Limit Theorem. For example, the gaps L_i of the Uniform process are equidistributed, but their sum $L_1+L_2+...+L_{n+1}$ is the length of the interval, which we know with certainty.

Statistical Measurements

Suppose we make n measurements $X_1,X_2,...,X_n$ of the same quantity. Implicitly we are assuming that these measurements are equidistributed and independent random variables. Each measurement has a distribution whose mean is the quantity we wish to measure. But the measurements are imperfect and so tend to be spread to a certain extent on both sides of the mean. Statisticians refer to this situation as a "random sample."

Definition. A random sample of size n is a set of n independent, equidistributed random variables $X_1,X_2,...,X_n$.

In the next two sections we will consider the problem of measuring the mean of a distribution using random samples. In particular we would like to know how small a random sample is sufficient for a given measurement. If we wish to determine the average number of cigarettes smoked per day by Americans, it would be highly impractical to ask every American for this information. Statistics enables one to make accurate measurements based on surprisingly small samples.

In addition to the measurement problem, we will also consider the problem of using a random sample as a means for making predictions of the future. The prediction will, of course, be a probabilistic one: with a certain probability the next measurement will lie within a certain range. For all the statistical problems we will study, we will assume only that the variance of each measurement X_i is finite. In most cases this is a reasonable assumption especially if the measurements lie in a finite interval. For example, the number of cigarettes smoked by one individual in one day is necessarily between 0 and 10^6.

The general procedure can be summed up in the following rule.

<ins>Main</ins> <ins>Rule</ins> <ins>of</ins> <ins>Statistics</ins>. In any statistical measurement we may assume that the individual measurements are distributed according to the normal distribution $N(m,\sigma^2)$.

To use this rule we first find the mean m and variance σ^2 from information given in our problem or by using the sample mean $\overline{m}$ and/or sample variance $\overline{\sigma}^2$ defined below. We then standardize the random variables required in the problem. Finally we use tables of the standard normal distribution to solve the problem. We will see many examples of this basic procedure. As stated, the main rule says only that our results will be "reasonable" if we assume that the measurements are normally distributed. We can actually assert more. In the absence of better information, we <u>must</u> assume that a measurement is normally distributed. In other words if several models are possible, we must use the normal model unless there is a significant reason for rejecting it.

When the mean m and/or the variance σ^2 of the measurements X_i are not known, the following random variables may be used as approximations.

The <u>sample</u> <u>mean</u> $\overline{m} = (X_1+X_2+\ldots+X_n)/n$ approximates m.

The <u>sample</u> <u>variance</u> $\overline{\sigma}^2 = \dfrac{(X_1-\overline{m})^2+(X_2-\overline{m})^2+\ldots+(X_n-\overline{m})^2}{n-1}$

approximates the variance σ^2.

For example an exam graded on a scale of 0-100 is given in a class of 100 students. The sample mean is found to be 81 with sample variance 100 (standard deviation 10). Based on this data, we can predict that if the exam is given to another student, the student will score between 61 and 100 with probability 0.95 (within $2\overline{\sigma}$ of $\overline{m}$). In actual exam situations, the distribution of an individual exam score is more complicated than the normal distribution, but in the absence of any better information we follow the Main Rule.

When the mean m is known but the variance is not, there is a slightly better approximation to the variance:

The sample variance (when m is known)

$$\bar{\sigma}^2 = \frac{(X_1-m)^2+(X_2-m)^2+\ldots+(X_n-m)^2}{n}$$

also approximates the variance σ^2.

The reason for the different denominators in the two expressions is subtle. We leave it as an exercise to show that the expectations of the random variables are

$$E(\bar{m}) = m \text{ and } E(\bar{\sigma}^2) = \sigma^2 \quad ,$$

where second equation holds for either sample variance. The distributions of the random variables $\bar{m}$ and $\bar{\sigma}^2$ are very important in statistics and we will undertake to compute some cases, leaving the rest as exercises.

4. Significance Levels

Let us begin with an example. We are presented with a coin having an unknown bias p. We are told that the coin is fair, but we are suspicious and would like to check this assertion. So we start tossing the coin. After 100 tosses we get only 41 heads. Do we have reason to suspect that the coin is not fair?

In such an experiment, we carefully examine the model we have postulated in order to determine what kind of behavior is consistent with the model. If the observed behavior is consistent with the model we have no reason to suppose that the coin is unfair. In this case the postulated model is the Bernoulli process with bias $p = 1/2$.

The average value of S_{100} is 50 for our postulated model. We are interested in the possible deviation of S_{100} from its mean value 50, because very large deviations are unlikely in the model but would not be if the coin is unfair. The usual statistical procedure in such a case is to determine precisely how large a deviation from the mean is reasonable in the model. Since S_{100} has the binomial distribution, we could in principle do this using only the formula for this distribution. However the computation is extremely difficult. On the other hand, we know that S_{100} is very close to having the normal distribution $N(50,25)$. In other words, $(S_{100}-50)/\ 5$ has approximately the standard normal distribution.

We now look in a table of the standard normal distribution. There we find that

$$P(-1.96 \leq \frac{S_{100}-50}{5} \leq 1.96) = 0.95,$$

or

$$P(40.2 \leq S_{100} \leq 59.8) = 0.95 .$$

Since 41 falls in this range, we conclude that our suspicions about the unfairness of the coin are groundless. The difference 1-0.95 = 0.05 is called the <u>significance level</u> of our test. We then say "the experiment has no significance at the 0.05 level." Notice that we say <u>no</u> significance. Statistically speaking, a significant result occurs only when a postulated model is <u>rejected</u>.

Looking at the reasoning a bit more carefully, we have said the following. Assuming that the coin is fair, about 95% of the time we will get between 40 and 60 heads when we toss the coin 100 times. But 5% of the time we will not be within this range.

> The significance level represents the probability that we will reject the postulated model even though this model is correct.

Notice the indirectness of this kind of reasoning. We say nothing about whether or not the coin is <u>really</u> fair or unfair, or even that it is fair or unfair with a certain probability. Statistics never tells one anything for certain, even in the weak sense of probabilistic certainty. All we can do is devise tests for determining at some significance level whether or not the data we have collected are <u>consistent</u> with the model. Because of the abbreviated terminology that statisticians and scientists frequently use when discussing the result of an experiment, one should be careful not to ascribe properties to statistical statements, which they do not possess.

The 0.05 significance level is so commonly used by statisticians and scientists that this level is assumed when no significance level is specified. The 0.01 significance level is also common, and an experiment is said to be very significant if this level is being used. For example, in our coin tossing test we found that getting 41 heads was not (statistically) significant. On the other hand, getting 39 heads would be significant but would not be very significant, while getting 35 heads would be very significant.

It is' important to point out that the choice of a significance level is part of the design of one's experiment. It cannot be "calculated" after the data are collected. Doing so is intellectual and scientific dishonesty of the worst kind, for if one does this consistently it violates the whole statistical framework within which the scientific community works. Generally speaking, the choice of a significance level is determined by considerations having nothing to do with probability or statistics. For example, if one is testing to see if a certain commonly used chemical could be a cause of a disease, we would certainly want a very significant result before recommending that the chemical be banned, with all the political and economic repercussions that such a decision could have.

Let us consider another example. We are given a die, and we wish to test whether it is loaded. We decide to consider whether "3" is special, and we choose to work at the 0.05 significance level. Our experiment consists of rolling the die 120 times, and we find that "3" comes up 25 times. Our postulated model is now the Bernoulli process with bias $p = 1/6$. The mean and variance of a single roll are $m = p = 1/6$ and $\sigma^2 = pq = 5/36$. Therefore the number of threes, S_{120}, is approximately $N(120/6,\ 120 \cdot 5/36) = N(20, 100/6)$, and hence $(S_{120}-20)\sqrt{6}/10$ is approximately $N(0,1)$. Our experiment is significant at the 0.05 level only if $|(S_{120}-20)\sqrt{6}/10| > 1.96$. In our case $S_{120} = 25$ so $|(S_{120}-20)\sqrt{6}/10| = |5\sqrt{6}/10| = \sqrt{3/2}$. This is not larger than 1.96. Therefore the experiment is not significant, and we have no reason to suspect that the die is loaded.

Rule of Thumb

A quick rule of thumb for testing the Bernoulli process (to be used only if one is in a hurry) is the following. If one tosses n times a coin with bias p, then the result is

significant if the number of heads lies outside $np \pm 2\sqrt{npq}$

very significant " " " " " " " $np \pm 3\sqrt{npq}$

5. Confidence Intervals

The concept of a confidence interval is a variation on the statistical themes we have just been describing. Instead of testing a hypothesis, one is interested in accuracy of a measurement or in prediction of the future.

Let us consider a very simple example of the prediction of the future. Suppose we have two competing airlines on a given route, both having the same departure time. Suppose that every day exactly 1000 passengers show up and that each one chooses one or the other airline with probability 1/2, independently of the other passengers. Both airlines want to be able to accommodate as many passengers as possible. They could do this, of course, by providing 1000 available seats. Needless to say this would be disasterously expensive, particularly since the probability that all 1000 seats would ever be needed is essentially zero. By providing 1000 seats we would have absolute certainty that there will never be an overflow; if we are willing to accept a 5% chance of an overflow, the number of seats we must provide decreases dramatically.

To compute this we again use the normal approximation of the binomial distribution. The model we are using is the Bernoulli process with bias $p = 1/2$. The variance of a single toss is 1/4. The number of passengers choosing one particular airline is S_{1000}, which has approximately the distribution N(500,250). Hence $(S_{1000}-500)/5\sqrt{10}$ is almost N(0,1). We now look up in a table of the standard normal distribution that number t for which

$$P(Y \leq t) = 0.95$$

We find that t = 1.645. This tells us that

$$P((S_{1000}-500)/5\sqrt{10} \leq 1.645) = 0.95$$

or $$P(S_{1000} \leq 526) = 0.95.$$

We need only provide 526 seats to have 95% <u>confidence</u> of not having an overflow. This is quite a dramatic drop from 1000 seats. Even for 99% confidence we need only a few more seats:

$$P((S_{1000}-500)/5\sqrt{10} \leq 2.33) = 0.99$$

or $$P(S_{1000} \leq 537) = 0.99.$$

We speak of the interval [0,526] as being a 95% *confidence interval* for S_{1000}. In general any interval [a,b] for which $P(a\leq X\leq b) = .95$ is called a 95% confidence interval for the random variable X. When X is normally distributed (or approximately normally distributed) with distribution $N(m,\sigma^2)$, we generally use either a *one-sided confidence interval* or a *two-sided confidence interval*. A one-sided interval is of the form $(-\infty,t]$ or of the form $[t,\infty)$. A two-sided interval is chosen to be symmetric about the mean: $[m-t,m+t]$. When testing statistical hypotheses, one uses either a one-sided or a two-sided confidence interval. The corresponding tests are then referred to as a *single-tail* or a *double-tail test* respectively.

Now we consider the problem of the accuracy of statistical measurements. Suppose we wish to determine the percentage of adult Americans who smoke. To find out this number we randomly sample n persons. How many persons do we have to sample in order to determine the percentage of smokers to two decimal place accuracy? Of course, we can determine this percentage to this accuracy with absolute certainty only by asking virtually the whole population, because there is always the chance that those not asked will all be smokers.

Therefore we must choose a confidence level. The usual level is 95% so we will use this.

The model we are using is the Bernoulli process with bias p, where p is the percentage we are trying to compute. Each person we ask will be a smoker with probability p. If we randomly ask n persons, the number who smoke divided by n will be an approximation $\bar{p}$ to p. This number is the sample mean $\bar{m} = S_n/n$. We use the Central Limit Theorem in its second manifestation. We find that $\bar{p}=\bar{m}$ has approximately the distribution $N(p,\sigma^2/n)$, where $\sigma^2 = pq$. Therefore $(\bar{m}-p)/(\sigma/\sqrt{n})$ is approximately $N(0,1)$. We require a two-sided interval in this problem:

$$P(-1.96 \leq (\bar{m}-p)/(\sigma/\sqrt{n}) \leq 1.96) = 0.95$$

or
$$P(|\bar{m}-p| \leq 1.96\sigma/\sqrt{n}) = 0.95$$

We want to choose n so that $|\bar{m}-p| \leq 0.005$ in order to have two place accuracy. That is, $1.96\sigma/\sqrt{n} = 0.005$ or $n \approx (1.54 \times 10^5)\sigma^2$. Unfortunately to compute $\sigma^2=pq$ we must know p. However we know that σ^2 takes its largest value when $p=q=1/2$. Therefore $n \leq (1.54\times10^5)(0.25) \simeq 3.85\times10^4$. In other words, to determine the percentage of smokers with 95% confidence we must sample up to 38,500 persons.

In practice one would first determine p to one decimal place accuracy. This requires only a sample of 385 persons. Using this number, one can compute σ more precisely. Using this better value of σ we can determine more precisely how many persons must be sampled in order to find p to two decimal place accuracy. For example suppose that with the smaller sample we find that $p = 0.65 \pm 0.05$. The worst case for σ^2 is now $pq=(0.6)(0.4) = 0.24$. We must then sample $n=37{,}000$ persons to determine p to within 0.005.

One must be careful not to confuse the accuracy with the confidence. The accuracy tells us how accurately we think we have measured a certain quantity. The confidence tells us the probability that we are right. To illustrate the distinction between these two concepts we consider the above measurement problem with two accuracies and three confidence levels. In general, improving confidence does not require much more effort while increasing accuracy requires a great deal of additional effort.

		Accuracy	
		0.05 (one decimal place)	0.005 (two decimal places)
Confidence	95%	385	38,500
	99%	667	66,700
	99.9%	1,089	108,900

The number of individuals that must be sampled to determine the percentage having a certain property (in the worst case p=1/2)

In the exercises we consider more examples of significance levels and confidence intervals. Some of these have a distinct air of the supernatural about them. How for example is it possible to make conclusions about the television preferences of a population of 200 million persons based on a sample of only 400 of them? In fact the size of the population is irrelevant to the statistical analysis. (It arises only when one confronts the problem of making a random sample from a very large population. This is a very difficult problem for statisticians.)

6.* The Proof of the Central Limit Theorem

If you are familiar with the concept of the Fourier transform, the proof of the Central Limit Theorem is not very difficult to understand. We will sketch the proof leaving the details as an exercise.

Let $X_1, X_2, \ldots$ be a sequence of independent equidistributed random variables having finite variance. Without loss of generality, we may assume that they are standard. Set $S_n = X_1 + X_2 + \ldots + X_n$. We wish to show that the distribution of $S_n/\sqrt{n}$ tends to the standard normal distribution.

Recall that the Laplace transform of a function $f(x)$ is defined to be the function $\phi(\lambda) = \int_0^\infty e^{-\lambda x} f(x)dx$, defined for $\lambda \geq 0$ when $f(x)$ is the density of a random variable. If we replace the nonnegative parameter λ by a purely imaginary one, $i\zeta$, for $-\infty < \zeta < \infty$, we obtain a transform known as the Fourier transform:

$$\psi(\zeta) = \int_{-\infty}^{\infty} e^{i\zeta x} f(x)dx,$$

defined for all real numbers ζ. By deMoivre's theorem,

$e^{i\zeta x} = \cos(\zeta x) + i\,\sin(\zeta x)$, so that $\psi(\zeta)$ may be written as

$$\psi(\zeta) = \int_{-\infty}^{\infty} \cos(\zeta x) f(x)dx + i\int_{-\infty}^{\infty} \sin(\zeta x) f(x)dx,$$

where each of the two integrals are real. When $f(x)$ is the density of a random variable X, we say $\psi(\zeta)$ is the characteristic function of X.

We begin by calculating some values of the characteristic function $\psi(\zeta)$. At zero we get

$$\psi(0) = \int_{-\infty}^{\infty} e^{0} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1 \ ,$$

since $f(x)$ is a density. Similarly, the values of the derivatives of $\psi(\zeta)$ at zero can be computed by "differentiation under the integral sign."

$$\psi^{(n)}(\zeta) = \int_{-\infty}^{\infty} \frac{d^n}{d\zeta^n}\, e^{i\zeta x} f(x)dx$$

$$= \int_{-\infty}^{\infty} (ix)^n e^{i\zeta x} f(x)dx,$$

so that $$\psi^{(n)}(0) = (i)^n \int_{-\infty}^{\infty} x^n e^{i\zeta x} f(x)dx$$

$$= (i)^n E(X^n) .$$

In other words, $\psi^n(0)$ is $(i)^n$ times the n^{th} moment of X.

If we assume that X has finite variance, then $E(X)$ and $E(X^2)$ exist, and we may apply the Taylor expansion theorem to conclude that

$$\psi(\zeta) = \psi(0) + \psi'(0)\zeta + \tfrac{1}{2}\psi''(0)\zeta^2 + o(\zeta^2), \qquad \text{as} \quad \zeta \to 0.$$

If X is standard, then

$$\psi(\zeta) = 1 - \tfrac{1}{2}\zeta^2 + o(\zeta^2), \qquad \text{as} \quad \zeta \to 0.$$

The Fourier transform satisfies the same convolution property as the Laplace transform:

Fact. If $\psi_X(\zeta)$ and $\psi_Y(\zeta)$ are the characteristic functions of random variables X and Y and if X and Y are independent, then $\psi_{X+Y}(\zeta) = \psi_X(\zeta)\psi_Y(\zeta)$ is the characteristic function of $X + Y$.

Proof. We may write the characteristic function $\psi_X(\zeta)$ of X as

$$\psi_X(\zeta) = \int_{-\infty}^{\infty} e^{i\zeta x} f(x)dx = E(e^{i\zeta X}).$$

By the multiplicative property of expectations of independent R.V.'s,

$$\psi_{X+Y}(\zeta) = E(e^{i\zeta(X+Y)}) = E(e^{i\zeta X}e^{i\zeta Y}) = E(e^{i\zeta X})E(e^{i\zeta Y})$$

$$= \psi_X(\zeta)\psi_Y(\zeta). \qquad \text{Q.E.D.}$$

Therefore, if $X_1, X_2, \ldots$ is a sequence of independent equidistributed standard random variables, having characteristic function $\psi(\zeta)$, then their sum S_n has characteristic function $\psi(\zeta)^n$. By a change of variables, the random variable $S_n/\sqrt{n}$ has characteristic function $\psi(\zeta/\sqrt{n})^n$. Utilizing the Taylor expansion computed above, we find that

$$\psi(\zeta/\sqrt{n})^n = (1 - \tfrac{1}{2}(\tfrac{\zeta}{\sqrt{n}})^2 + o(\tfrac{\zeta^2}{n}))^n \qquad \text{as} \quad \frac{\zeta}{\sqrt{n}} \to 0$$

$$= (1 - \frac{1}{2n}\zeta^2 + o(\tfrac{1}{n}))^n \qquad \text{as} \quad n \to \infty$$

$$\text{(with } \zeta \text{ fixed)}$$

Now we know from calculus that

$$(1 - \frac{\zeta^2/2}{n})^n \to e^{-\zeta^2/2}, \qquad \text{as} \quad n \to \infty,$$

and we leave it as an exercise to show that this also works when we have the extra $o(\frac{1}{n})$ term. Therefore, the characteristic function of the standardized sum $S_n/\sqrt{n}$ approaches $e^{-\zeta^2/2}$ as $n \to \infty$, for every fixed ζ.

We now suspect that $e^{-\zeta^2/2}$ is the characteristic function of the standard normal distribution. This can be proved a number of ways. One could first show that the convolution of normal distributions is normal so that if $X_1, X_2, \ldots$ are all standard normal distributions then so is $S_n/\sqrt{n}$. It then follows by the above result that the characteristic function of the standard normal distribution must be $e^{-\zeta^2/2}$. One can also compute this characteristic function directly by differentiating under the integral sign and using an integration by parts. We leave this as an exercise.

The Central Limit Theorem follows from the above calculation and the following two properties of the characteristic function:

Property 1 (Fourier inversion). *Different probability distributions have different characteristic functions*.

Indeed, if $\psi(\zeta)$ has the property that $\int_{-\infty}^{\infty}|\psi(\zeta)|d\zeta < \infty$, then one may use the *Fourier inversion formula* to compute $f(x)$ in terms of $\psi(\zeta)$:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\zeta x}\psi(\zeta)d\zeta .$$

Property 2 (Continuity). *If a sequence of characteristic functions* $\psi_1, \psi_2, \ldots$ *converges to a characteristic function* ψ *in the sense that for all* ζ,

$$\lim_{n\to\infty} \psi_n(\zeta) = \psi(\zeta) ,$$

then the probability distributions corresponding to the $\psi_n(\zeta)$ *converge to that of* $\psi(\zeta)$ *as* $n \to \infty$.

7^*. The Law of Large Numbers

The law of large numbers is the statement that is often taken as justification of the definition of probability in terms of frequency. For example, what does it mean to say that the probability is 1/2 for getting a head when a fair coin is tossed? In the frequentist point of view, one says that this means the proportion of heads in a very large number of tosses will be very close to 1/2. But this is really begging the question in some sense as we will see.

Let $X_1, X_2, \ldots$ be independent equidistributed random variables with common mean m and common variance $\sigma^2 < \infty$. We would like to say that $(X_1+X_2+\ldots+X_n)/n$ approaches m as $n \to \infty$. But these are random variables so we can only speak of the probability that the limit is m.

The Law of Large Numbers

$$P(\lim_{n\to\infty} \frac{X_1+X_2+\ldots+X_n}{n} = m) = 1.$$

This is essentially just a psychological theorem, for it does not provide the information necessary for concrete applications. The Central Limit Theorem is far more useful, and in fact the law of large numbers is a consequence of the Central Limit Theorem. We leave the proof as an exercise.

In any case the law of large numbers is a purely mathematical theorem. In order for it to make sense we must already have the concepts of probability, random variables, means, variances, etc. We cannot use this as the definition of probability. But we cannot even use the law of large numbers as a justification of the frequentist point of view. This point of view says that probabilities represent a physically measureable quantity (at least in principle). But there is no concept of a physical "measurement" corresponding to the mathematical concept of the limit

$$\lim_{n\to\infty} \frac{X_1+X_2+\ldots+X_n}{n}$$

The relationship between physical experiments and the theory of probability is much more subtle than the frequentist point of view would have one believe.

The law of large numbers is not very useful in applications because it does not specify how large a sample is required to achieve a given accuracy. However it does have interesting theoretical applications. We will see one in section VII.2 (the Shannon Coding Theorem). Another theorem which has great usefulness in probability theory is the Bienaymé-Chebyshev Inequality. Its importance stems primarily from its simplicity.

Bienaymé-Chebyshev Inequality Let X be a random variable with mean $E(X) = m$ and variance $Var(X) = \sigma^2$, then for all $t > 0$,

$$P(|X - m| \geq t) \leq \sigma^2/t^2.$$

Proof

Suppose that X is a continuous R.V. with density $f(x)$. The proof in the case of an integer R.V. is similar. Clearly we may assume that m is zero, for if not we just replace X by $X - m$.

$$\sigma^2 = \text{Var}(X) = E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx \geq \int_{|x|\geq t} x^2 f(x)dx \geq \int_{|x|\geq t} t^2 f(x)dx$$

$= t^2 \int_{|x|\geq t} f(x)dx = t^2\ P(\,|X| \geq t)$. The second inequality is a consequence of the fact that $x^2 \geq t^2$ in the domain of integration. If we now solve for $P(\,|X| \geq t)$ we get the desired inequality.

The last result we will consider in this section is one of the most astonishing facts about probability: the Kolmogorov Zero-One law. As with the other theorems in this section it has little practical usefulness, but it has many theoretical applications. The law of large numbers, for example, can be proved using it.

Suppose that $X_1, X_2, \ldots$ is a sequence of random variables which are independent but not necessarily equidistributed. A _tail event_ A is an event such that

(1) A is defined in terms of the random variables $X_1, X_2, \ldots$

(2) A is independent of any finite set of the X_i's, i.e.

$P(A|\,(X_1=t_1)\cap(X_2=t_2)\cap\ldots\cap(X_n=t_n)) = P(A)$,

for any $n<\infty$ and any set of t_i's.

Kolomogorov Zero-One Law If A is a tail event, then $P(A)=0$ or 1.

At first it seems that there cannot be any tail events except for Ω and Ω^c, because tail events seem both to depend on the X_i's and not to depend on the X_i's. However there are, in fact, many nontrivial examples. Here is one. Toss a fair coin infinitely often, and write $X'_n = \begin{cases} +1 \text{ if the } n^{th} \text{ toss is heads} \\ -1 \text{ if the } n^{th} \text{ toss is tails} \end{cases}$

Now let A be the event "$\sum_{n=1}^{\infty} \frac{X'_n}{n}$ converges." This is a tail event because the converge or divergence of a series is determined by the terms of the series but is independent of any finite set of them. We all should know at least two examples from calculus: $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges but $\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$ converges to $\ell n(2)$. What we are doing is to change the signs of the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$ randomly and independently. P(A) is the probability that a random choice of signs yields a convergent series. The zero-one law tells us that P(A) can only be 0 or 1; there are no other possibilities. In fact P(A) = 1; we leave this as an exercise.

As another example, suppose that a monkey is trained to hit the keys of a typewriter and does so at random, each key having a certain probability of being struck each time, independently of all other times. Let A be the event "the monkey eventually types out Shakespeare's _Hamlet._" Again this is clearly a tail event and so P(A) = 0 or 1. This is easy to see. _Hamlet_ has about 2×10^5 characters and could be written with a typewriter having 100 keys. Suppose each key has probability .01 of being typed. The probability of typing _Hamlet_ is $p = (.01)^{2 \times 10^5}$ during a given "session" of 2×10^5 keystrokes. The probability of _not_ typing _Hamlet_ in one session is $q = 1-p<1$. The probability that in _infinitely_ many sessions the monkey _never_ types out _Hamlet_ is $\lim_{n \to \infty} q^n = 0$.

Therefore P(A) = 1. On the other hand, the expected waiting time until the monkey types out Hamlet is about $10^{400,000}$ keystrokes. If the monkey could type one keystroke every nanosecond, the expected waiting time until the monkey types out Hamlet is so long that the estimated age of the universe is insignificant by comparison.

Needless to say this is not a practical method for writing plays. The Kolomogorov zero-one law has little practical usefulness. But it does have theoretical uses, and it shows how counter-intuitive probability theory can be.

Normal Distribution Function

Values of $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2}\, dx = P(X \le x)$, where X is normally distributed with mean 0 and variance 1

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

8. Exercises for

Chapter IV Statistics and the Normal Distribution

Variance

1. Suppose that X is a random variable whose density is

$$\text{dens}(X=x) = \begin{cases} (\beta-1)x^{-\beta} & \text{if } x > 1 \\ 0 & \text{if } x \le 1 \end{cases}, \text{ where } \beta > 1. \text{ Show:}$$

(a) X has neither mean nor variance if $1 < \beta \le 2$.

(b) X has mean $\frac{\beta-1}{\beta-2}$ if $\beta > 2$, but has no variance if $2 < \beta \le 3$.

(c) X has variance $\dfrac{\beta-1}{(\beta-2)^2(\beta-3)}$ if $\beta > 3$.

2. Let X be the random variable S'_n in the symmetric Bernoulli process random walk model. Let $Y = X^2$. Then X and Y are obviously dependent random variables. Show that $\text{Cov}(X,Y) = 0$. Hence the converse to the Fact in section IV.1 is false.

3. Verify the following formula, which holds for arbitrary random variables $X_1, X_2, \ldots, X_n$ (not necessarily independent) so long as both sides exist:

$$\text{Var}(X_1+X_2+\ldots+X_n) = \sum_i \text{Var}(X_i) + 2 \sum_{i<j} \text{Cov}(X_i, X_j).$$

Use this formula and the exchangeability of the gaps in the uniform process to compute $\text{Var}(X_{(k)})$.

4. Prove that if X and Y are independent continuous R.V.'s, then $E(XY) = E(X)E(Y)$. To do this one must split X into positive R.V.'s X^+ and X^- such that $X = X^+ - X^-$. For example, define X^+ by
$$X^+ = \begin{cases} X & \text{if } X \geq 0 \\ 0 & \text{if } X < 0. \end{cases}$$
Do the same for Y. Note also that $\text{dens}(XY = z \mid X = x) \neq \text{dens}(Y = \frac{z}{x} \mid X = x)$.

5. Prove that for any two random variables X and Y,

$E(XY) \leq \sqrt{E(X^2)\,E(Y^2)}$ (Schwartz Inequality). Use this to prove that the correlation coefficient of X and Y satisfies $|\rho(X,Y)| \leq 1$. Also show that $\rho(X,Y) = \dfrac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)}$.

6. Let $X_1, X_2, \ldots, X_n$ be a set of independent random variables not necessarily equidistributed but all having the same mean and variance σ^2. Prove that the sample mean has expectation m and that both sample variances have expectation σ^2. Also compute the variance of the sample mean. Can you compute the variance of either sample variance?

Notice that the denominator must be different for the two sample variances. This denominator is called the number of degrees of freedom by statisticians. Intuitively, each estimation of a parameter of the unknown distribution causes a loss of one degree of freedom in the random sample.

7.* In exercise III.10 we saw that if X and Y are independent then $g(X)$ and $h(Y)$ are also independent for any two functions g and h. It follows that $g(X)$ and $h(Y)$ are also uncorrelated. Prove the converse: if $g(X)$ and $h(Y)$ are uncorrelated for <u>every</u> pair of functions g and h, then X and Y are independent.

Normal Distribution

8. Using the normal distribution table compute:

(a) $P(-.5 \le X \le .5)$, where X is $N(0.1)$ [Answer: 0.383]

(b) $P(X \le -2)$, where X is $N(0,1)$ [" 0.0228]

(c) $P(Y \ge 5)$, where Y is $N(0,4)$ [" 0.0062]

(d) $P(1 \le Y \le 4)$, where Y is $N(-2,9)$ [" 0.1359]

9. Find a number α such that

(a) $P(X \ge \alpha) = 0.03$, where X is $N(0,1)$ [1.88]

(b) $P(-\alpha \le X \le \alpha) = 0.08$, where X is $N(0,1)$ [1.555]

(c) $P(-2-\alpha \le Y \le \alpha-2) = 0.10$, where Y is $N(-2,9)$ [4.935]

(d) $P(Y \ge \alpha) = 0.98$, where Y is $N(0,4)$ [-4.126]

10. Show explicitly that the normal distribution $N(m,\sigma^2)$ really does have variance σ^2.

Significance Levels

11. A prestigious scientific journal announces as part of its editorial policy that only results significant at the .01 level will be acceptable for publication (and conversely any result significant at the .01 level is acceptable). They reason that by doing so tneir readership will have the confidence that at most 1% of the published results will be incorrect. Discuss the fallacy

of this policy. [Hearing about this new policy, 1000 conscientious experimenters formulate 1000 wrong scientific hypotheses. On the average 10 of them would find a significant result, and these 10 would then be entitled to publish their results. Let's say that these 10 articles constitute the first issue of the journal after the new policy is instituted. We would find that the journal policy allowed 100% of the published results to be wrong. Clearly the journal policy is a result of a misunderstanding of the nature of statistical hypothesis testing: significant at the .01 level does not mean that there is only a 1% chance that one is wrong.]

12. In a scientific paper you read the following: "In four of our experiments the data are significant at the .05 level. The fifth experiment, however, is significant at the .01 level!" What is misleading about this?

13. A population scientist believes that roughly 50% of the population is female, but doesn't want to be too hasty. So he decides to be cautious and to test whether or not at least 45% of the population is female. To do this he takes a random sample of 100 persons. If he discovers that only 40 of them are female, does he have sufficient evidence to reject the model that (at least) 45% of the population is female? Use a Bernoulli process with $p = .45$ and a one-sided significance test.

14. A statistician wonders just how careful the scientist in exercise 13 was when he made his random sample. Is 40 significantly different from the expected value of 50? Is 40 significantly smaller than 50? What do you think of the sampling technique of the scientist? [Answer: Yes; yes; not much.]

15. You own a company that produces medium quality left-handed screws. About 1% of the screws produced by one machine are defective. As the screws are produced, the defective ones are found and discarded. A count is kept of the number of defective screws produced each hour. The machine is readjusted whenever the number of defective screws produced is significantly greater than 1%. You may regard this as a Bernoulli process. The machine makes 10,000 screws per hour. Describe a procedure for determining when the machine is out of adjustment at the 0.05 level and at the 0.01 level.

16. A congressman wishes to vote according to the "will of the people" on a certain bill. Now in this case one wishes to know whether the percentage p of his constituency in favor of the bill is above or below 50%. Clearly if p is close to 1/2 a rather large sample will be required to distinguish between the two possibilities. How is this reflected in a statistical test? For example suppose that a poll is made soliciting the opinion of a certain number of voters chosen at random. Use a .05 significance level to decide what the congressman should do in each of the following cases.

Number of pollees	Number of pollees in favor of the bill
100	54
100	41
500	267
500	225
1000	534

Note that the congressman has three choices in each case: (a) vote for the bill, (b) vote against the bill or (c) order a larger sample be taken.

17. A company wishes to test the effectiveness of a new magazine advertising campaign. It decides that the campaign is effective if the proportion of subscribers to the magazine who use their product is twice as large as the proportion of non-subscribers who do so. A 10% significance level is agreed upon. It is known that 15% of the general population use the firm's product. A sample of 50 subscribers is ordered and it is found that 10 use the product. What does this suggest about the advertising campaign? [Answer: One cannot say that the advertising campaign was unsuccessful].

18. A study has shown that in a certain profession the women are receiving only 88% as much on the average, as their male counterparts receive. However, the study is several years old and a women's organization wants to determine if the women in this profession are losing relative to the men. It is known that the men in the profession now receive 138% of the pay they received when the above study was conducted. A random sample of female professionals is made. The average pay of these women was found to be the following (as a percentage of the current average pay of men in the profession): 70%, 78%, 80%, 83%, 84%, 86%, 87%, 96%.

Compute the sample mean and sample standard deviation. At the .05 significance level are the women in this profession losing ground relative to the men? Does this coincide with your "gut feeling" in this problem? [Answer: No; no]

19. The Food and Drug Administration (FDA) suspects that a drug company is producing a certain pill with a purity less than that required by law. The law allows at most 5 parts per million (ppm) of a certain impurity. An FDA laboratory tests a random sample of 50 pills. They find that the pills have a sample mean impurity of 5.4 ppm and a sample variance of 4.38 $(\text{ppm})^2$. Can they assert that the pills do not comply with the law? We will return to this question in exercise V.47. [Answer: If the law only requires that the average amount of impurity is 5ppm then we cannot reject the possibility that the manufacturer is complying with the law.]

20. A paper company was a major polluter of a small river for many years. When antipollution laws were enacted it reacted slowly at first but later made a major effort to control its pollution. Unfortunately the firm suffered from its earlier recalcitrance by acquiring a public image as a major polluter. Indeed a very large sample revealed that close to 90% considered the firm to be a major polluter. To counter this they began a public relations campaign. After the campaign a random sample of 200 individuals were asked whether the company was still a major polluter. It was found that 174 felt this way. Did the campaign have a significant effect? [Answer: No]

21. (Hans Zeisel). Dr. Benjamin Spock, author of a famous book on baby care, and others were initially convicted of conspiracy in connection with the draft during the Vietnam war. The defense appealed, one ground being the sex composition of the jury panel. The jury itself had no women, but chance and challenges could make that happen. Although the defense might have claimed that the jury lists (from which the jurors are chosen) should contain 55% women, as in the general population, they did not. Instead they complained that six judges in the court averaged 29% women in their jury lists, but the seventh judge, before whom Spock was tried, had fewer, not just on this occasion but systematically. The last 9 jury lists for that judge contained the following counts:

	Women	Men	Total	Proportion women
	8	42	50	0.16
	9	41	50	0.18
	7	43	50	0.14
	3	50	53	0.06
	9	41	50	0.18
	19	110	129	0.15
	11	59	70	0.16
	9	91	100	0.09
	11	34	45	0.24
Grand totals	86	511	597	0.144

Did the jury lists for this judge have a significantly smaller percentage of women? Because of the seriousness of the case, use an extremely small significance level: 0.0001.

22. It has been said that Chevalier de Méré actually observed the subtle distinction in probability between obtaining at least one six in four throws of a die and obtaining at least one double-six in twenty-four throws of a pair of dice. See exercise II.28

Could he have done so? Since he did not have access to the elaborate machinery of the normal distribution and significance tests, it is difficult to imagine what he might have deduced about any observations he might have made.

However, we could ask what is the probability that he would not have observed a difference between the two experiments in a certain number of trials. Let X_n be the number of times out of n trials that at least one six is obtained in four throws of a die. Let Y_n be the corresponding random variable for the double-six trial. What is the probability that $X_n - Y_n$ is positive? Since de Méré's calculation showed that Y_n should have been the more probable of the two, such an observation would have shocked him. For definiteness compute this probability for $n = 10$, 100, 200, 500, 1000, 2000 and 5000. How many times would de Méré have had to have tried both possibilities in order to reject (at the 5% level) this explanation of his perplexity? How many throws of one or two dice does this involve? What can one one conclude? [Answer: 3900 and 109,200. Conclusion: Either de Méré tried this experiment a great number of times or else we cannot dismiss the possibility that he did not in fact succeed in detecting the difference between the two probabilities.]

23. A political scientist wishes to determine if there is a significant difference between the preferences of voters in two similar neighborhoods of a city with respect to an upcoming race for mayor. Samples of 30 voters are taken from each of the neighborhoods. In one sample 12 voters prefer the incumbent while in the other neighborhood 19 do so. Using a 10% significance level,

decide whether there is a significant difference between the neighborhoods. [Answer: Yes]

24. A medical researcher samples 100 records of adults having diagnosed coronary heart disease from one city, taking care to ensure that the sample is random. The average cholesterol value for these individuals was found to be 296, and the sample standard deviation was 30. The researcher then took a random sample of 200 people from the same city who never had diagnosed heart disease. The mean cholesterol value for this sample was 310, and the sample standard deviation was 50. Do individuals without diagnosed heart disease have a significantly larger cholesterol value than those with diagnosed heart disease? [Answer: Yes]

25. A small college soccer team won its conference championship 9 times in the first 20 years of its existence. Then for the next twenty years it won only 3 times. Is this significant? Is it very significant? [Answer: Yes; no].

26. When the president of the company in exercise 20 discovered that the questionnaire used in the post-campaign sample included the word "still," she was incensed: the question seemed biased in favor of a yes answer. Accordingly, she immediately proceeded to write her own questionnaire and take a new random sample. The only change was the omission of the word "still." In this new sample of 200 individuals only 160 felt that the company was a major polluter. Did the alteration of the questionnaire have a significant effect? Did the public relations campaign have a significant effect? [Answers: Yes; yes]

Confidence Intervals

27. Using the informati(n in exercise 24, give a 90% confidence interval for the following:

(a) the individual cholesterol values of individuals without heart disease;

(b) the individual cholesterol values of indicividuals with heart disease;

(c) the mean cholesterol value of all individuals without heart disease;

(d) the mean cholesterol value of all individuals with heart disease.

[Answers: 296 ± 49.35; 310 ± 82.25; 296 ± 5; 310 ± 6]

28. If a set of grades on a statistics examination are approximately normally distributed with a mean of 82 and a standard deviation of 6.9, find:

(a) The lowest passing grade if the lowest 10% of the students are given F's.

(b) The highest B if the top 5% of the students are given A's.

29. The average life of a certain type of engine is 10 years, with a standard deviation of 3.5 years. The manufacturer replaces free all engines tnat tail while under guarantee. If he is willing to replace only 2% of the engines that fail, how long a guarantee should he offer? Assume a normal distribution.

30. The braking distances of two cars, F and C, from 50 k.p.h. are normally distributed, one with mean 30m and standard deviation 8m, the other with mean 35m and standard deviation 5m. If they both approach each other on a narrow mountain road and first see each other when they are 100m apart, what is the probability that they avoid a collision? [Answer: .9999]

31. If the probability of a male birth is 0.512, what is the probability that there will be fewer boys than girls in 1000 births? [Answer: 0.215]

32. A multiple-choice quiz has 100 questions each with four possible answers of which only one is the correct answer. What is the probability that sheer guesswork yields from 10 to 30 correct answers for 40 of the 100 problems about which the student has no knowledge?

33. A firm wishes to estimate (with a maximum error of 0.05 and a 98% confidence level) the proportion of consumers who use its product. How large a sample will be required in order to make such an estimate if the preliminary sales reports indicate that about 25 percent of all consumers use the firm's product? How large a sample would be needed if no preliminary information were available?

34. A sponsor of a weekly television program is interested in estimating the proportion of the city population who regularly watch its program. The sponsor wishes the estimate to be made with a 90% confidence level and an error of at most 4%. The sponsor has no information concerning the proportion of viewers who

watch the program. How large a sample will be required to make the estimate?

35. A person has just hired a building contractor to build a house. The house will be built in three stages. First, the contractor lays the foundation; second, the frame and exterior are built; and last a subcontractor puts in the wiring, plumbing, and interior. Each stage must be completed before the next is started. In attempting to get an estimate of when the house will be totally completed, the purchaser is able to get the following information from those in charge of each stage.

Stage	Expected Time of Completion of Stage (in Weeks)	Standard Deviation (in Weeks)
I	3	1
II	8	2
III	5	2

What is the expected value and standard deviation of completion time for the house, assuming the completion times of the stages are independent?

[Answer: 16 weeks and 3 weeks]

36. The College Entrance Examination Board verbal and mathematical aptitude scores are approximately normally distributed with mean 500 and standard deviation 100 except that scores above 800 and and below 200 are arbitrarily reported as 800 and 200 respectively. What percentage of the students taking the verbal exam score above

800 or below 200? [Answer: about 0.3%]

37. You are the head of a polling company, and you have a contract to determine the percentage of the electorate in favor of a candidate. There are 1,000,000 members of the electorate, and each member chooses his/her opinion independently. Your contract specifies that you must determine the percentage to within 1% with 5% confidence or to within 5% with 1% confidence. Which is cheaper?

38. A professor at a small college walks to school each day. On the average the trip takes 15 minutes with a standard deviation of 3 minutes. Assume a normal distribution. If the professor's first class is at 10:30 AM, when must the professor leave home in order to be 95% certain of arriving on time? If the college serves coffee from 10:00 AM to 10:30 AM how often would the professor have coffee before class if the professor left home at 10:10 AM every day? [Answers: 10:10 AM; 0.952].

39. Suppose that resisters can be purchased each with a resistance that is uniformly distributed between $900\ \Omega$ and $1100\ \Omega$. If 10 such resisters are connected in parallel, what is the probability that their total resistance will be within 5% of $10,000\ \Omega$?

40. When a thumbtack is tossed, it falls on its flat head with probability p. What must you do to find p to within $p/10$ at significance level 0.05? at significance level 0.01?

41. You own a telephone company that services two cities A and B, each having 5000 customers. You would like to link your exchange with the more distant city, C. You estimate that

during the busiest time each customer will require a line to C with probability .01. You want to be sure that there are enough lines to C so that there is only a 1% chance that at the busiest time some customer will be unable to get a line to C. Each trunkline to C will cost $10,000. You have two options. Either link A and B as if they were separate exchanges or link the entire exchange to C. In the second option, additional equipment costing $50,000 would be needed. Which option is cheaper?

42. A clinical trial is conducted to determine if a certain type of drug has an effect on the incidence of a certain disease. A sample of 100 rats was kept in a controlled environment and 50 of the rats were given the drug. Of the group not given the drug (the control group), there were 12 incidences of the disease, while 9 of the other group contracted it. Compute a 90% confidence interval for the difference in probability of contracting the disease between a rat given the drug and a control rat. [Answer: 0.06 ± 0.134]

Hypothesis Testing

43. A statistician named Burr relates the following story.

> "Having bought a bag of roasted chestnuts, the author walked home in the dark eating them with much gusto. After eating about 20, he arrived home, and, in opening the remaining 10 under the light, he found that 7 contained worms. What is the probability that none of the 20 contained worms? Or to phrase the problem

better for statistical analysis: If there were only 7 wormy chestnuts among the original 30, what is the probability of drawing the first 20 all free from worms?"

44. Since there is no reason to believe that the salaries of individuals will be normally distributed, only in very large samples can we expect the mean to be normally distributed. With this in mind reexamine exercise 18. Regardless of the distribution of salaries of the female professionals, half the salaries will be above the median salary. Suppose that the salaries of the male professionals are known to have a median that is 97% of the mean. This situation would be typical since the presense of a few very high salaries can cause the mean to be somewhat unrepresentative. Does it now appear that the female professionals are gaining or losing relative to the men? [Answer: the median salary of the women is 83.5%, while 88% of 97% is 85.4%. So it appears that they are losing ground, but the result is again not significant. Using a Bernoulli process to model this, if the median salary of all women were 85.4%, then each woman's salary would have probability 0.5 of exceeding this figure. We observe that 3 of the 8 salaries do so. The probability that 3 (or fewer) would do so by chance alone is $(0.5)^8 + 8(0.5)^8 + \binom{8}{2}(0.5)^8 + \binom{8}{3}(0.5)^8 = 0.363$ which is too large for us to reject this model.]

45. (Certificate) In a certain survey of the work of chemical research workers, it was found, on the basis of extensive data, that on average each one required no fume hood for 60 per cent

of the time, one for 30 per cent and two for 10 percent; three or more were never required. If a group of four chemists worked independently of one another, how many fume hoods should be available in order to provide adequate facilities for at least 95 per cent of the time?

Compute the probability distribution of the number of fume hoods needed by the four chemists. Then use this to answer the question.

46. In exercise 45, how many fume hoods would be required to satisfy a group of 50 chemists at least 95 per cent of the time? Use a normal approximation.

47. A governmental agency is responsible for protecting the fish populations of the lakes in a certain region. By means of many observations in the past it sets lower bounds for populations, in each lake, of various species of fish. If it is later found that one species in a given lake has gone below the specified lower bound, the agency has the power to enforce limits on the pollutants which the factories bordering on the lake may discharge into the lake. How can the agency determine whether the lower limit has been reached? One way to do so is to employ the procedure described in exercise II.23. Let the lower limit be L. We first capture n fish, tag them and return them to the lake. Some time later, we drop a net, capture m fish and count how many are tagged. Describe how to find a number T so that there is only a 5% chance that T or more tagged fish will be found when

there are L or more fish in the lake. We will return to this problem in exercise VI. XX.

48. (Silvey) An investigation was carried out on two suggested antidotes to the consequences of drinking, these being (a) 2 lb of mashed potatoes and (b) a pint of milk. Ten volunteers were used, five to each antidote, the allocation to antidote being random. One hour after each had drunk the same quantity of alcohol and swallowed the appropriate antidote, a blood test was carried out and the following levels (mg/ml) of alcohol in the blood were recorded:

(a)	76	52	92	80	70
(b)	110	96	74	105	125

By means of the runs test, decide whether there is sufficient evidence to conclude that one treatment is more effective than the other.

49. (Guenther) Suppose that it is hypothesized that twice as many automobile accidents resulting in deaths occur on Saturday and Sunday as on other days of the week. That is, the probability that such accidents occur on Saturday is 2/9, on Sunday is 2/9, and on each other day of the week is 1/9. From the national record file, cards for 90 accidents are selected at random. These yield the following distribution of accidents according to the days of the week:

Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
30	6	8	11	7	10	18

Do these data tend to support or contradict the hypothesis? Use a 5% significance level.

50. We wish to test whether or not the successive outcomes of a roulette wheel are random. For simplicity we will only record whether the ball fell into a red or a black slot of the wheel. In twenty spins of the wheel we observe the sequence: RRBRRBBBBRBRRRRRBBBR. Applying the runs test and using a 5% significance level, are the successive outcomes random? What does this suggest about this roulette wheel?

51. (Pazer & Swanson) A political scientist wishes to determine if the political preference of homeowners is independent of their immediately adjacent neighbors. A sequence of sixteen homeowners, along the same side of a street, were interviewed, and based upon their responses were designated as either more conservative than their median C, or less conservative than their median L. Here is the resulting sequence:

L, C, L, C, C, C, C, L, L, C, L, L, L, L, C, C.

Using the run test and a 5% level of significance, determine whether there is any evidence that political opinions are independent of one's neighbors (at least for this particular street).

52.* We all have taken laboratory courses at some time or other, and the temptation to "fudge" data on our report has certainly occurred to us. What we may not have realized is that one can devise a statistical test to determine whether or not such fudging took place. Suppose that a biologist wishes to prove that a certain genetic trait follows the classical Mendelian laws. In this theory a trait is determined by two genes, one acquired from each of the two parents. Let us say that there are two alleles (possibilities) for a given gene, one dominant A and one recessive a . Then there are three different genotypes: AA, Aa and aa. Let us suppose, as it often happens, that AA and Aa are indistinguishable. By successive inbreeding the biologist has access to two individuals known to have genotypes AA and aa, respectively. When these are crossed the offspring all have genotype Aa. But when two of the offspring are crossed we find that the three genotypes AA, Aa and aa appear among their offspring with probabilities 1/4, 1/2 and 1/4 respectively. Of course since we cannot actually distinguish AA from Aa, this means that on the average 3/4 of the offspring exhibit the dominant trait and 1/4 exhibit the recessive one. Let us suppose that the biologist produces 10,326 offspring from a pair of Aa parents. He observes that 7746 have the dominant trait and 2580 the recessive one. These are very close to the expected numbers 7744.5 and 2581.5 so he concludes that the experiment tends to support the hypothesis that this trait obeys the Mendelian laws.

Compute the probability that such an experiment would actually result in as close a fit with the theory as the biologist actually found. At the 5% level can one reject the hypothesis that the experiment was properly carried out? [Answers: about 3.6%, yes].

The Law of Large Numbers

53. How many times must one toss a fair coin in order to have 95% confidence that it really is fair? Compare the number obtained by using the Bienaymé-Chebychev inequality with what we get using the Central limit theorem.

54. Let X be a standard random variable (i.e. $E(X) = 0$ and $E(X^2) = 1$). Using the Bienaymé-Chebychev inequality, compute the smallest α so that:

(a) $P(-\alpha \leq X \leq \alpha) \geq .95$

(b) $P(-\alpha \leq X \leq \alpha) \geq .99$

(c) $P(X > \alpha) \leq .05$

(d) $P(X > \alpha) \leq .01$

Compare these values with the corresponding ones for case of X being $N(0,1)$.

55*. Let X be a nonnegative random variable. Prove that $P(X \geq \alpha) \leq E(X)/\alpha$, for any $\alpha > 0$, whether X has a variance or not. This is known as Markov's inequality. Show that Markov's inequality implies the Bienaymé-Chebychev inequality.

56*. Prove the Law of Large Numbers for probability distributions having finite variance, using the Central Limit Theorem. The Law of Large Numbers is in fact true for all probability distributions possessing a mean, but his is much more difficult to prove.

57*. Show that $\sum_{n=1}^{\infty} X_n'/n$ converges with probability 1, where X_n' is equally likely to be either $+1$ or -1.

58*. Explore experimentally what it means for a random variable not to have an expectation. Write computer programs to simulate the St. Petersburg game (Exercise III.29) and the gangster distribution (exercise III.36). In each case print out two columns of numbers. The first column shows the number of times the random experiment has been repeated, and the second shows the sample average of all the trials made so far. For the gangster distribution one should print a third column showing the median of all the trials made so far. This last number is the hardest to compute, and unlike the sample mean it requires that all the previous trials be stored in an array. Give an intuitive interpretation for what it means for a random variable not to have a finite expectation.

Chapter V Conditional Probability

The theory of probability consists largely in making precise the probabilistic language that already forms part of our language. In effect the purpose of this course is to learn to "speak probability" properly. The lowest level of our probabilistic language is the event. This corresponds to simple phrases that are either true or false. For example in the Bernoulli process H_i is the event "the i^{th} toss is heads". Random variables represent the next level: simple quantitative questions. For example one might ask: "how long must one toss a coin until the first head appears?" If we use the convention 0 = false and 1 = true, every event may also be regarded as a random variable by using the indicator.

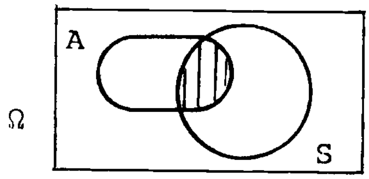
Conditional probability allows probabilistic reasoning. That is, we may now ask compound questions. For example, "if the first toss of a coin is tails, how long must one wait until the first head appears?" Moreover, we can split apart and combine such questions into new questions. The precise meaning of such expressions is not always obvious and is the source of many seeming paradoxes and fallacies. As a simple example, the question, "if the first toss is heads, is the second toss heads?" is very different from the question, "are the first two tosses heads?" The probability of the first is p while that of the second is p^2.

1. Discrete Conditional Probability

We begin with the definition and properties of the conditional probability of events.

Definition. Let A and S be events such that $P(S) > 0$. The conditional probability of A given S is

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



The event S is called the condition. The conditional probability $P(A|S)$ answers the question: "if S has occured, how probable is A?" In effect we have altered our sample space. Since we know that S has occured, the sample space is now S. The event A given that S has occured must now be interpreted as $A\cap S$, and the probability is $P(A\cap S)$ normalized by the probability of S so that the total probability is 1. Ordinary probabilities are the special case of conditional probabilities where the condition is the sample space Ω: $P(A) = P(A|\Omega)$.

Law of Alternatives

Suppose that instead of knowing that a certain event has occurred, we know that one of several possibilities has occurred, which are mutually exclusive. Call these alternatives $A_1, A_2, \ldots$. There may possibly be infinitely many alternatives. More precisely the A_i form a set of alternatives if

(1) $A_i \cap A_j = \emptyset$ if $i \neq j$ (mutually exclusive)

(2) $\cup_i A_i = \Omega$ (exhaustive)

(3) $P(A_i) > 0$ for all i .

Then for any event B:

$$\boxed{P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots}$$

Law of Alternatives

To verify this law we simply expand and cancel. The A_i are disjoint so the events $B \cap A_i$ are also disjoint.

$$P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots$$

$$= \frac{P(B \cap A_1)}{P(A_1)} P(A_1) + \frac{P(B \cap A_2)}{P(A_2)} P(A_2) + \ldots$$

$$= P(B \cap A_1) + P(B \cap A_2) + \ldots$$

$$= P((B \cap A_1) \cup (B \cap A_2) \cup \ldots)$$

$$= P(B \cap (A_1 \cup A_2 \cup \ldots))$$

$$= P(B \cap \Omega)$$

$$= P(B)$$

If the alternatives A_i are not exhaustive, we can still make sense of the law of alternatives provided all probabilities involved are conditioned by the event $A = \cup_i A_i$. More precisely we shall call a set of events A_i a <u>set of alternatives for</u> A if

(1) $A_i \cap A_j \neq 0$ if $i \neq j$ (mutually exclusive)

(2) $\cup_i A_i = A$

(3) $P(A_i) > 0$ for all i

Then for any event B:

$$\boxed{P(B|A) = P(B|A_1)P(A_1|A) + P(B|A_2)P(A_2|A) + \ldots}$$

Conditional Law of Alternatives

Bayes' Law

One of the features of probability as we have developed it so far is that all events are treated alike: in principle no events are singled out as "causes" while others become "effects." Bayes' law, however, is traditionally stated in terms of causes and effects. Although we will do so also, one should be careful not to ascribe metaphysical significance to these terms. Historically, this law has been misapplied in a great number of cases precisely because of such a misunderstanding.

We are concerned with the following situation. Suppose we have a set of alternatives A_1, A_2,... which we will refer to as "causes". Suppose we also have an event B which we will call the "effect". The idea is that we can observe whether the effect B has or has not occurred but not which of the causes $A_1, A_2, \ldots$ has occurred. The question is to determine the probability that a given cause occurred given that we have observed the effect. We assume that we know the probability for each of the causes to occur, $P(A_i)$, as well as the conditional probability for B to occur given each cause, $P(B|A_i)$. The probability $P(A_i)$ is called the <u>a priori</u> probability of A_i, and we seek the probability $P(A_i|B)$ which we call the <u>a posteriori</u> probability of A_i. If the alternatives A_i represent various experimental hypotheses, and B is the result of some experiment, then Bayes' law allows us to compute how the observation of B changes the probabilities of these hypotheses.

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

Bayes' Law

Proof of Bayes' Law

Let A and B be any two events having positive probability. By the definition of conditional probability,

$$P(B|A) = \frac{P(AB)}{P(A)} \qquad \text{and} \qquad P(A|B) = \frac{P(AB)}{P(B)}.$$

As a result we have two ways to express P(AB):

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

Solving for $P(A|B)$ gives:

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

Now apply this fact to the case for which A is A_i and use the law of alternatives to compute the denominator. The resulting expression is Bayes' law.

Law of Successive Conditioning

Suppose we have n events $B_1, B_2, \ldots, B_n$ such that $P(B_2 \wedge B_3 \wedge \ldots \wedge B_n) > 0$. Then we can compute $P(B_1 \wedge B_2 \wedge \ldots \wedge B_n)$ using a sequence of conditional probabilities.

$$P(B_1 \wedge B_2 \wedge \ldots \wedge B_n) =$$

$$P(B_1 | B_2 \wedge \ldots \wedge B_n) \cdot P(B_2 | B_3 \wedge \ldots \wedge B_n) \ldots P(B_{n-1} | B_n) \cdot P(B_n)$$

Law of Successive Conditioning

To prove this we just expand and cancel. This law corresponds to the intuitive idea that the probability of several events occurring is the product of their individual probabilities. This idea is correct provided we interpret "individual probability" to mean the appropriate conditional probability.

By using the law of alternatives and the law of successive conditioning we split the computation of an ordinary or a conditional probability into a succession of conditional probabilities. In effect we form compound, nested, conditional questions out of simple questions.

Independence

Suppose that A and B are two events. If either A or B has probability zero of occurring, then A and B are trivially independent events. If $P(A)>0$ and $P(B)>0$, then the concept of the independence of the events A and B is best stated by using conditional probability. Namely each of the following are equivalent statements:

(1) A and B are independent

(2) $P(A|B) = P(A)$

(3) $P(B|A) = P(B)$

Using this terminology we can see much more clearly that the independence of two events A and B means that knowing one has occurred does not alter the probability that the other will occur, or equivalently that the measurement of one does not affect the measurement of the other.

2. Gaps and Runs in the Bernoulli Process

Recall that T_i is the gap between the $(i-1)^{st}$ and the i^{th} success. We claimed that the T_i are independent random variables, using an intuitive probabilistic argument. We now have the terminology for making this argument rigorous. The key notion is the law of alternatives.

Consider the conditional probability of the gap T_{i+1} being n given that all of the preceding gaps are known:

$$P(T_{i+1} = n \mid (T_1=k_1)\wedge(T_2=k_2)\wedge\ldots\wedge(T_i=k_i)).$$

Computing this probability is quite easy for it corresponds to exactly two patterns of H's and T's (up to some toss):

$$\overbrace{\underbrace{TT\ldots TH}_{k_1}\ \underbrace{TT\ldots TH}_{k_2}\ TT\ldots\ldots\ldots\ H\underbrace{TT\ldots TH}_{k_i}}^{k=k_1+k_2+\ldots+k_i}$$

$$P((T_1=k_1)\wedge(T_2=k_2)\wedge\ldots\wedge(T_i=k_i)) = q^{k-i}p^i$$

$$\underbrace{TT\ldots TH}_{k_1}\ \underbrace{TT\ldots TH}_{k_2}\ TT\ldots\ldots\ldots\ H\underbrace{TT\ldots TH}_{k_i}\ \underbrace{TT\ldots TH}_{n}$$

$$P((T_1=k_1)\wedge(T_2=k_2)\wedge\ldots\wedge(T_i=k_i)\wedge(T_{i+1}=n)) = q^{k+n-i-1}\ p^{i+1}$$

Therefore

$$P(T_{i+1}=n \mid (T_1=k_1) \wedge (T_2=k_2) \wedge \ldots \wedge (T_i=k_i)) = \frac{q^{k+n-i-1}p^{i+1}}{q^{k-i}p^{i}} = q^{n-1}p$$

Although the above computation is not very difficult, there is an easier way to see it. Think of the condition

$$(T_1=k_1) \wedge \ldots \wedge (T_i=k_i)$$

as changing our sample space. The new sample space consists of all infinite sequences of H's and T's, but renumbered starting with $k+1 = k_1+k_2+\ldots+k_i+1$. This new sample space is identical to the Bernoulli sample space except for the renumbering and the fact that T_{i+1} is now the waiting time for the first success. Therefore

$$P(T_{i+1}=n \mid (T_1=k_1) \wedge (T_2=k_2) \wedge \ldots \wedge (T_i=k_i)) = q^{n-1}p.$$

The key to the effective use of conditional probability is that it changes the sample space and hence the interpretation of the random variables defined on the old sample space.

We now apply the law of alternatives. The events

$$(T_1=k_1)\wedge(T_2=k_2)\wedge\ldots\wedge(T_i=k_i) \quad ,$$

as the k_j's take on all positive integer values, form a set of alternatives; for the set of sample points belonging to none of them is an event whose probability is zero.

$P(T_{i+1}=n)$ =

$$= \sum_{k_1}\ldots\sum_{k_i} P(T_{i+1}=n|(T_1=k_1)\wedge\ldots\wedge(T_i=k_i))\cdot P((T_1=k_1)\wedge\ldots\wedge(T_i=k_i))$$

$$= \sum_{k_1}\cdots\sum_{k_i} q^{n-1}p \; P((T_1=k_1)\wedge\ldots\wedge(T_i=k_i))$$

$$= q^{n-1}p \sum_{k_1}\cdots\sum_{k_i} P((T_1=k_1)\wedge\ldots\wedge(T_i=k_i))$$

$$= q^{n-1}p.$$

Notice that the only fact we used about the events $(T_1=k_1)\wedge\ldots\wedge(T_i=k_i)$ was that they form a set of alternatives. An immediate consequence is that the gaps T_i are all equidistributed. Furthermore, if we use the definition of conditional probability we have that

$$P(T_{i+1}=n) = P(T_{i+1}=n \mid (T_1=k_1)\wedge \ldots \wedge(T_i=k_i))$$

$$= \frac{P((T_1=k_1)\cap \ldots \wedge(T_i=k_1)\wedge(T_{i+1}=n))}{P((T_1=k_1)\cap \ldots \wedge(T_i=k_i))}$$

or

$$P((T_1=k_1)\wedge \ldots \wedge(T_i=k_i)\cap(T_{i+1}=n)) = P((T_1=k_1)\wedge \ldots \wedge(T_i=k_i))P(T_{i+1}=n)$$

By mathematical induction we have that

$$P((T_1=k_1)\cap \ldots \cap(T_i=k_i)\cap(T_{i+1}=n)) = P(T_1=k_1)\ldots P(T_i=k_i)P(T_{i+1}=n),$$

i.e. that the T_i are independent.

We can now see more clearly how the T_i are related. They have the same distribution, but they are not the same. They have this property because the measurement of the i^{th} gap "really" occurs in a different sample space than the first gap, but this new sample space is identical to the ordinary Bernoulli sample space except for how we number the tosses.

We went into detail for this argument to illustrate a nontrivial use of the law of alternatives. We will be more abbreviated in the future.

As an illustration of the law of alternatives, we consider a problem mentioned in chapter II. Namely, what is the probability that a run of h heads occurs before a run of t tails? Let A be this event. We solve this problem by using the following fact: when a run of less than h heads is "broken" by getting a tail, we must "start over" and similarly for runs of tails.

First we use the law of alternatives:

$$P(A) = P(A|X_1 = 1)P(X_1 = 1) + P(A|X_1 = 0)P(X_1 = 0) .$$

Write $u = P(A|X_1 = 1)$ and $v = P(A|X_1 = 0)$ so that

$$P(A) = up + vq.$$

Next we use the conditional law of alternatives for each of $P(A|X_1=1)$ and $P(A|X_1=0)$. Consider the first one. We know that we got a head on the first toss so the run has started. We then "wait" to see if the run will be broken. That is, let B_t be the waiting time for the first tail starting with the second toss. Either we get a tail and break the run or we get enough heads so that A occurs. More precisely,

$$P(A|X_1=1) = P(A|(X_1=1)\cap(B_t<h))P(B_t<h)$$

$$+ P(A|(X_1=1)\cap(B_t\geq h))P(B_t\geq h) .$$

For the first alternative the run of heads has been broken by a tail. Hence $P(A|(X_1=1)\wedge(B_t<h)) = P(A|X_1=0) = v$, for the earlier heads have no effect on subsequent tosses. All that matters is that we "started" with a tail. On the other hand, $P(A|(X_1=1)\wedge(B_t\geq h)) = 1$ because the condition implies that A has in fact occurred. Therefore,

$$\begin{aligned} u = P(A|X_1=1) &= vP(B_t<h) + 1\cdot P(B_t\geq h) \\ &= v(1-p^{h-1}) + 1\cdot(p^{h-1}) \end{aligned}$$

Remember that for B_t we start counting on the second toss, so that $P(B_t<h)$ is really the conditional probability $P(B_t<h|X_1=1)$.

The computation for $P(A|X_1=0)$ is analogous to that above. Let B_h be the waiting time for the first head starting with the second toss. Then

$$\begin{aligned} P(A|X_1=0) &= P(A|(X_1=0)\wedge(B_h<t))P(B_h<t) \\ &\quad + P(A|(X_1=0)\wedge(B_h\geq t))P(B_n\geq t). \end{aligned}$$

Here $P(A|(X_1=0)\wedge(B_n\geq t)) = 0$ because the condition implies that A has <u>not</u> occurred. The probability $P(A|(X_1=0)\wedge(B_h<t))$ is u because the run of tails has been broken. Therefore,

$$v = P(A|X_1=0) = u\left(1-q^{t-1}\right) + 0\cdot\left(q^{t-1}\right) .$$

Combining the two equations above gives us the system of equations:

$$u = v\cdot(1-p^{h-1}) + p^{h-1}$$

$$v = u\cdot(1-q^{t-1}).$$

Solve for u and v and substitute:

$$P(A) = up + vq = \frac{p^{h-1}(1-q^{t})}{p^{h-1}+q^{t-1}-p^{h-1}q^{t-1}}$$

We check this by considering the special case h=t and p=q=1/2. As we expect by symmetry, P(A) = 1/2.

3. Sequential Sampling

In most sampling situations, for example sampling people in a population, we generally sample the population without replacement, i.e. the same individual cannot be chosen more than once in one sample. For such a sampling procedure, the successive choices are not independent of one another; for with each choice, the population (and hence the sample space) gets smaller.

For very large populations this would seem to be a small effect. But on smaller populations it can be pronounced. For example, suppose we play a card guessing game. We draw a card at random from a deck, try to guess the suit, look to see if our guess was correct and then place the card aside. If we continue to sample the cards this way, the probabilities for getting a card of a given suit change constantly. Indeed, we will always know for certain what the suit of the last card drawn will be.

The problem of sequential sampling is to describe the dependence of each of the choices on the other choices. The idealized model is the following. We have an urn containing r red balls and b black balls. We select a ball at random from the urn, note its color and then place it aside. This procedure is repeated until n balls have been chosen from the urn. Define the random variables X_i by:

$$X_i = \begin{cases} 1 \text{ if the } i^{th} \text{ ball is red} \\ 0 \text{ if the } i^{th} \text{ ball is black} \end{cases}$$

The problem is to find the distributions and the correlations of the X_i.

Notice that we have switched the roles of the balls and the boxes. In the occupancy model we place balls into boxes. In the sampling model the balls become the positions in the sample and the boxes become individuals in a population. In the sequential sampling model it is traditional to view the population more concretely as a collection of colored balls in an urn.

Consider the first choice X_1. The probability distribution of X_1 is

$$p_0 = P(X_1 = 0) = \frac{b}{r+b}$$

$$p_1 = P(X_1 = 1) = \frac{r}{r+b}$$

Next consider the second choice X_2. To compute its distribution we must use the law of alternatives. For example $P(X_2=0|X_1=0)$ is $\frac{b-1}{r+b-1}$ because there is one fewer black ball in the urn.

$$p_0 = P(X_2=0) = P(X_2=0|X_1=0)P(X_1=0)+P(X_2=0|X_1=1)\cdot P(X_1=1)$$

$$= \frac{b-1}{r+b-1} \cdot \frac{b}{r+b} + \frac{b}{r+b-1} \cdot \frac{r}{r+b}$$

$$= \frac{(b-1)\cdot b+b\cdot r}{(r+b-1)(r+b)}$$

$$= \frac{b(r+b-1)}{(r+b-1)(r+b)}$$

$$= \frac{b}{r+b}$$

Similarly,

$$p_1 = P(X_2=1) = \frac{r}{r+b} .$$

The random variables X_1 and X_2 are equidistributed! This is quite unexpected. One wonders whether this is an accident of algebra or there is some deep principle here. If the latter, we would expect that all the X_i have the same distribution. As we shall see this is indeed the case.

The seeming paradox arises from the fact that we are not considering the random variables conditionally. For example, in the card guessing game above, if we chose not to look at the first 51 cards sampled, we would have no reason to suppose that the last card sampled has any special properties: it doesn't "know" that the other cards have been sampled.

Exchangeability

As often happens in mathematics, the situation only becomes clear when we consider it from a broader perspective. Consider the joint distribution of all the X_i's:

$$C_{i_1,\ldots,i_n} = P((X_1=i_1)\cap(X_2=i_2)\cap\ldots\cap(X_n=i_n))$$

where the $i_1,\ldots,i_n$ take on the values 0 and 1 arbitrarily. We compute this by using the law of successive conditioning:

$$C_{i_1,\ldots,i_n} = P(X_1=i_1)P(X_2=i_2|X_1=i_1)P(X_3=i_3|(X_1=i_1)\cap(X_2=i_2))\ldots$$

For example, if $n = 6$ and $(i_1,i_2,\ldots,i_6) = (0,1,1,0,1,1)$, then

$$C_{0,1,1,0,1,1} = \frac{b}{r+b} \cdot \frac{r}{r+b-1} \cdot \frac{r-1}{r+b-2} \cdot \frac{b-1}{r+b-3} \cdot \frac{r-2}{r+b-4} \cdot \frac{r-3}{r+b-5}$$

$$= \frac{(b)_2 (r)_4}{(r+b)_6}$$

Each factor is the number of balls of the appropriate color at the time divided by the number of balls in the urn at the time. More generally, if we have drawn a sequence of k reds and j blacks, then the probability is

$$C_{i_1,\ldots,i_n} = \frac{(b)_j (r)_k}{(r+b)_{j+k}}$$

The probability of drawing a given sequence of reds and blacks depends only on the number of reds and blacks drawn. In other words $P((X_1=i_1)\cap(X_2=i_2)\cap\ldots\cap(X_n=i_n))$ is the same if we permute the $i_1,\ldots,i_n$ leaving the X_j's alone (or equivalently if we permute the X_j's leaving the i_j's alone). For example, $P((X_1=i_1)\cap(X_2=i_2)\cap\ldots\cap(X_n=i_n))=P((X_1=i_2)\cap(X_2=i_1)\cap\ldots\cap(X_n=i_n))$

Since we can compute the individual distributions of the X_j's from the joint distribution by taking marginals, we immediately get that the X_j's are equidistributed. Moreover the joint distribution of, say, X_1 and X_5 is the same as that of X_1 and X_2:

$$P((X_1=i_1)\cap(X_5=i_2)) = P(X_1=i_1)\cap(X_2=i_2))$$

and the latter is easy to compute. In general the joint distribution of any k of the X_j's is the same as that of the first k of them. All these facts follow from the fact that the joint distribution of the X_j's is unchanged when we permute the X_j's. This is the real reason that the choices in sequential sampling are equidistributed.

Definition. Random variables (either integer or continuous) $X_1,\ldots,X_n$ are said to be exchangeable when their joint distribution (or density) is a symmetric function.

An example of a set of exchangeable random variables we have already seen is a set of independent, equidistributed random variables. If X_1,X_2 and X_3 are independent, equidistributed integer R.V.'s, then

$$\begin{aligned} P((X_1=i_1)\cap(X_2=i_2)\cap(X_3=i_3)) &= P(X_1=i_1)P(X_2=i_2)P(X_3=i_3) \\ &= p_{i_1}p_{i_2}p_{i_3} \end{aligned}$$

But as we have just seen, being exchangeable is not as strong a condition as being independent and equidistributed.

We mention in passing that being exchangeable is not really that much more general than being independent and equidistributed. There is a deep theorem of probability theory which, roughly speaking, says that every set of exchangeable random variables can be "synthesized" from independent equidistributed random variables by suitable conditioning.

The Pólya Urn Process

A slightly more general sampling model than sampling either with replacement or without replacement is called the Pólya Urn Process. In this process we begin with an urn containing r red balls and b black balls. We draw a ball at random. If it is red, we put the drawn ball plus c more red balls into the urn. If it is black, we put the drawn ball plus d more black balls into the urn. We then repeat this. Sampling with replacement is the case c=d=0, and sampling without replacement is the case c=d=-1.

This process was originally introduced as a model of epidemics. If we think of the red balls as diseased individuals, then each discovery of a red ball increases the likelihood that other balls will be red (c>0). There are obvious defects in such a model which we will not pursue. We will just think of this process as a general form of sampling.

As before let $X_1,\ldots,X_n$ be the successive results of drawing n balls in the Pólya Urn Process. The computation of the joint distribution of the X_j's is much the same as before. For example,

$$C_{1,1,0,0,1} = P((X_1=1)\cap(X_2=1)\cap(X_3=0)\cap(X_4=0)\cap(X_5=1))$$

$$= \frac{r}{r+b}\cdot\frac{r+c}{r+b+c}\cdot\frac{b}{r+b+2c}\cdot\frac{b+d}{r+b+2c+d}\cdot\frac{r+2c}{r+b+2c+2d}$$

In general, the X_j's will not be exchangeable, but if c=d, they are. For those who like formulas, the probability of drawing j blacks and k reds in any order is

$$\frac{(\frac{r}{c})^{(j)}(\frac{b}{c})^{(k)}}{(\frac{r+b}{c})^{(j+k)}}$$

provided that d=c≠0 and is

$$\frac{r^jb^k}{(r+b)^{j+k}}$$

if c=d=0.

4.* The Arcsine Law of Random Walks

We use the same notation as in section III.1D. The arcsine law is the distribution of the time of the last visit of a random walk to the origin. More precisely, consider a random walk up to time 2n, and ask when the last time was that the random walk visited the origin. Let L_{2n} be the time of the last visit. Clearly the random walk can only visit the origin during even-numbered times. We want, therefore, to compute $P(L_{2n}=2k)$ for all k between 0 and n.

Now examine the event $(L_{2n}=2k)$. We can rephrase this event as saying that the random walk was at the origin at time 2k, and that from then on the random walk never visited the origin:

$$(L_{2n}=2k) = (S'_{2k}=0) \cap (S'_{2k+1} \neq 0) \cap (S'_{2k+2} \neq 0) \cap \ldots \cap (S'_{2n} \neq 0) .$$

The law of successive conditioning tells us that

$$P(L_{2n}=2k) = P(L_{2n}=2k \mid S'_{2k}=0) P(S'_{2k}=0) \quad .$$

Now $P(L_{2n}=2k \mid S'_{2k}=0)$ is the same as $P(L_{2n-2k}=0)$. This follows from the independence of the steps of the random walk. We know how to compute $P(S'_{2k}=0)$ so we must find a way to compute

$$P(L_{2n-2k}=0) = P((S_1'\neq 0) \cap (S_2'\neq 0) \cap \ldots \cap (S_{2n-2k}'\neq 0)).$$

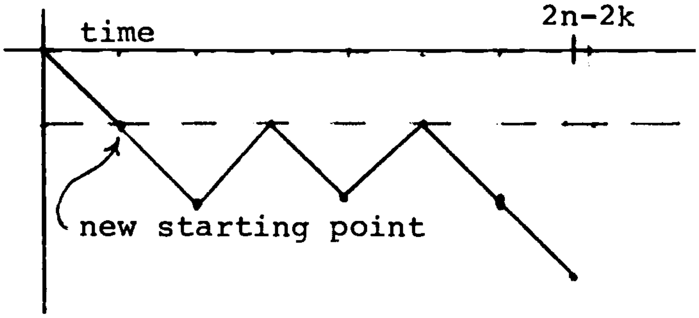
For this we use the law of alternatives, conditioning on which way the walk went during the first step:

$$\begin{aligned} P(L_{2n-2k}=0) &= P(L_{2n-2k}=0 \mid X_1'=+1) P(X_1'=+1) \\ &\quad + P(L_{2n-2k}=0 \mid X_1'=-1) P(X_1'=-1) \ . \\ &= \tfrac{1}{2}P(L_{2n-2k}=0 \mid X_1'=+1) + \tfrac{1}{2}P(L_{2n-2k}=0 \mid X_1'=-1). \end{aligned}$$

By symmetry both of the above conditional probabilities are the same. Thus

$$P(L_{2n-2k}=0) = P(L_{2n-2k}=0 \mid X_1' = -1).$$

If we now change coordinates we may consider the "walk" as starting at (1,-1):



We now see that we have a familiar situation. $P(L_{2n-2k}=0 \mid X_1'=-1)$ is the probability that the random walk travels no farther to the right than the origin in the first 2n-2k-1 steps, i.e. $P(M_{2n-2k-1}=0)$, where M_n is the maximum position of the random walk in the first n steps (see section IV.2a). Thus

$$P(L_{2n-2k}=0|X_1'=-1)=P(M_{2n-2k-1}=0)=p(2n-2k-1,0)+p(2n-2k-1,1).$$

It is easy to see that $p(2n-2k-1,0) = 0$. Thus

$$P(L_{2n-2k}=0) = p(2n-2k-1,1) = \binom{2n-2k-1}{n-k} \frac{1}{2^{2n-2k-1}} \cdot$$

$$= \frac{(2n-2k-1)!}{(n-k)!(n-k-1)!} \cdot \frac{1}{2^{2n-2k-1}}$$

$$= \frac{(2n-2k)!}{(n-k)!(n-k)!} \cdot \frac{(n-k)}{(2n-2k)} \cdot \frac{1}{2^{2n-2k-1}}$$

$$= \binom{2n-2k}{n-k} \frac{1}{2^{2n-2k}} = p(2n-2k,0).$$

Returning to our original problem, we find that

$$P(L_{2n}=2k) = P(L_{2n-2k}=0)P(S_{2k}'=0)$$

$$= p(2n-2k,0)p(2k,0).$$

We now show why this is called the arcsine law. Using Stirling's formula, we find that

$$p(2n-2k,0)p(2k,0) = \binom{2n-2k}{n-k}\binom{2k}{k}\frac{1}{2^{2n}}$$

$$= \frac{(2n-2k)!}{((n-k)!)^2}\;\frac{(2k)!}{(k!)^2}\;\frac{1}{2^{2n}}$$

$$\approx \frac{(2n-2k)^{2n-2k}\;\sqrt{2\pi(2n-2k)}}{(n-k)^{2n-2k}\;\;2\pi(n-k)}\;\frac{(2k)^{2k}}{k^{2k}}\;\frac{\sqrt{2\pi(2k)}}{2\pi k}\;\frac{1}{2^{2n}}$$

$$= \frac{2^{2n-2k}}{\sqrt{\pi(n-k)}}\;\frac{2^{2k}}{\sqrt{\pi}}\;\frac{1}{2^{2n}} = \frac{1}{\pi\sqrt{k(n-k)}}\;.$$

Hence

$$P(L_{2n}=2k) \approx \frac{1}{\pi\sqrt{k(n-k)}}\;.$$

Set $x = \frac{k}{n}$. Then $P(L_{2n}=2k) \approx \frac{1}{\pi n\sqrt{x(1-x)}}$.

Thus when n is large the distribution function of L_{2n}, $P(L_{2n} \leq 2k)$, is approximately equal to the area from 0 to k/n of the function $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}$:

$$P(L_{2n} \leq 2k) \approx \int_0^{k/n} \frac{dx}{\pi\sqrt{x(1-x)}}\;.$$

Using the substitution $y = \sqrt{x}$, we find that

$$\int \frac{dx}{\pi\sqrt{x(1-x)}} = \int \frac{2\sqrt{x}dy}{\pi\sqrt{x(1-x)}} = \int \frac{2}{\pi} \frac{dy}{\sqrt{1-y^2}}$$

$$= \frac{2}{\pi} \arcsin(y) = \frac{2}{\pi} \arcsin(\sqrt{x}).$$

Thus

$$P(L_{2n} \leq 2k) \approx \frac{2}{\pi} \arcsin(\sqrt{k/n}).$$

Summarizing,

Let L_{2n} be the time of the last visit of a 2n-step random walk to the origin. Then

$$P(L_{2n}=2k) = p(2n-2k,0)p(2k,0) \approx \frac{1}{\pi\sqrt{k(n-k)}},$$

$$\text{and } P(L_{2n} \leq 2k) \approx \frac{2}{\pi} \arcsin(\sqrt{k/n}).$$

The Arcsine Law

Here is an example of this law. A gambler plays a fair game, betting one dollar every ten seconds on the toss of a fair coin. If the gambler plays for a whole year, what is the probability that the last time the gambler "broke even" occurred after one day of play (i.e. the gambler had either a "winning streak" or a "losing streak" for 364 days). The arcsine law provides an excellent approximation of this probability:

$$P(L_{3153600} \leq 8640) \approx .0333,$$

i.e. about one chance in 30. This is amazingly large. One can analyze the fluctuations of coin tossing in even more detail. The surprising conclusion is that it is very unlikely for a random walk to spend close to the same amount of time on both sides of the origin. Thus while the average value of S'_n is zero for all n; nevertheless, individual random walks _with high probability_ will exhibit behavior that a naive observer would regard as being very _un_random.

5. Continuous Conditional Probability

Consider the Uniform process of sampling n points from the interval [0,a]. Suppose we know that the k^{th} point in order, $X_{(k)}$, was t . Given this information, what is the smallest point, $X_{(1)}$? It seems reasonable to answer this question with the conditional probability distribution

$$F(x) = P(X_{(1)} \leq x | X_{(k)} = t)$$

of $X_{(1)}$ given that $X_{(k)} = t$. Unfortunately, we know that $P(X_{(k)} = t) = 0$; so that, technically speaking, the above conditional probability does not make sense.

On the other hand, it is easy to compute what $P(X_{(1)} \leq x|X_{(k)} = t)$ <u>ought</u> to mean. For suppose that $X_{(k)} = t$. This means that exactly k-1 points have fallen in the interval [0,t]. The random variable $X_{(1)}$ should therefore be reinterpreted as the first order statistic of the Uniform process of dropping k-1 points in the interval [0,t]. Therefore

$$P(X_{(1)} \leq x|X_{(k)} = t) = 1 - (\frac{t-x}{t})^{k-1}$$

Notice that we do not have to "choose" the k-1 points which are to fall in [0,1]. This choice is already implicit in the fact that we have conditioned by $(X_{(k)} = t)$.

Although we cannot make sense, in general, of a conditional probability $P(A|B)$ when $P(B) = 0$, we can do so when B is the event $(X=t)$ for a continuous random variable X. We will call this the <u>continuous conditional probability</u> (although we shall often drop the adjective "continuous.") The following is the formal definition of this concept. But one rarely uses the definition directly. As with the ordinary conditional probability, the best way to compute a continuous conditional probability is to regard the condition as defining a new sample space and to reinterpret the events and random variables of the old sample space in this new sample space.

Definition. For an event A and a continuous random variable X, the continuous conditional probability of A given that X = t is

$$\begin{aligned} P(A|X=t) &= \lim_{\varepsilon\to 0} P(A|t<X\le t+\varepsilon) \\ &= \lim_{\varepsilon\to 0} \frac{P(A\cap(t<X\le t+\varepsilon))}{P(t<X\le t+\varepsilon)} \end{aligned}$$

provided that it exists.

Notice that we do not divide by ε. The reason is that ε appears in both the numerator and the denominator. If you wish, $P(A|X=t)$ is the limit:

$$\lim_{\varepsilon\to 0} \frac{P(A\cap(t<X\le t+\varepsilon))/\varepsilon}{P(t<X\le t+\varepsilon)/\varepsilon}$$

Both the numerator and the denominator in this limit have "densities" as their limits.

Just to make sure, we will compute $P(X_{(1)}\le x|X_{(k)}=t)$ directly from the definition to see that we get what we computed earlier. We know from our computation in section IV.6 that

$$P(t<X_{(k)}\le t+\varepsilon) = n\binom{n-1}{k-1}\frac{t^{k-1}\cdot\varepsilon\cdot(a-t-\varepsilon)^{n-k}}{a^n} + \frac{\varepsilon^2}{a^2}\cdot\text{(complicated expression)}$$

Next we compute $P((X_{(1)}>x)\cap(t<X_{(k)}\le t+\varepsilon))$. Except for a term having a factor of ε^2, this event corresponds to having k-1 points in the interval [x,t], one point in $[t_1 t+\varepsilon]$ and the rest in $[t+\varepsilon,a]$. Therefore,

$$P((X_{(1)}>x)\cap(t<X_{(k)}\leq t+\varepsilon)) =$$

$$n\binom{n-1}{k-1}\ \frac{(t-x)^{k-1}\cdot\varepsilon\cdot(a-t-\varepsilon)^{n-k}}{a^n} + \frac{\varepsilon^2}{a^2}\ .\ \text{(complicated expression)}$$

We now combine the above two computations.

$$\frac{P((X_{(1)}>x)\cap(t<X_{(k)}\leq t+\varepsilon))}{P(t<X_{(k)}\leq t+\varepsilon)} =$$

$$= \frac{n\binom{n-1}{k-1}(t-x)^{k-1}\cdot\varepsilon\cdot(a-t-\varepsilon)^{n-k}\cdot a^{-n} + \varepsilon^2\cdot\text{(expression)}}{n\binom{n-1}{k-1}t^{k-1}\cdot\varepsilon\cdot(a-t-\varepsilon)^{n-k}\cdot a^{-n} + \varepsilon^2\cdot\text{(expression)}}$$

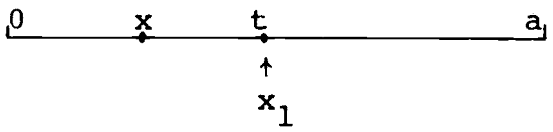
$$\longrightarrow \left(\frac{t-x}{t}\right)^{k-1} \qquad \text{as } \varepsilon\longrightarrow 0\ .$$

Finally we get $P(X_{(1)}\leq x|X_{(k)}=t) = 1 - (\frac{t-x}{t})^{k-1}$ as before. Needless to say this is the hard way to compute this.

Consider one more example. Suppose we know that the first point, X_1, is t. Given this, what is the smallest point, $X_{(1)}$? Again the answer is a probability distribution: $F(x) = P(X_{(1)}\leq x|X_1=t)$.

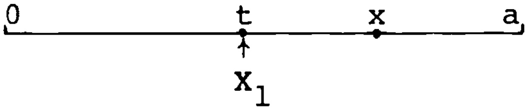
We split this into two cases.

Case 1 x<t.

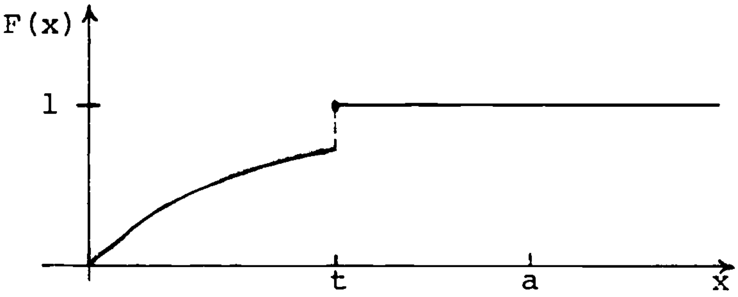


By the independence of the X_i's in the Uniform process, $P(X_{(1)} \leq x | X_1 = t)$ should be interpreted as $P(X_{(1)} \leq x)$ but in the Uniform process of sampling $\underline{n-1}$ points from $[0,a]$. That is, knowing that X_1 is t does not influence whether any other points are smaller than x. Therefore, $P(X_{(1)} \leq x | X_1 = t) = 1 - (\frac{a-x}{a})^{n-1}$.

Case 2 $x \geq t$



Here the fact that $X_1 = t$ means that $(X_{(1)} \leq x)$ has occurred. Therefore, $P(X_{(1)} \leq x | X_1 = t) = 1$.



The Conditional Distribution $F(x) = P(X_{(1)} \leq x | X_1 = t)$

Combining these two cases, we find that $P(X_{(1)} \leq x | X_1 = t)$ is not a continuous function. When we condition by $(X_1 = t)$, the random variable $X_{(1)}$ becomes discontinuous. This will often be the case for conditional distributions. Later we will develop techniques for using discontinuous random variables as if they were continuous.

The Continuous Law of Alternatives

One of the most important facts about continuous conditioning is that the law of alternatives has a continuous version. Indeed continuous conditional probabilities are important primarily because of this. Recall that if a set of events $A_1, A_2, \ldots$ form a set of alternatives, then the probability of any event B is

$$P(B) = \sum_i P(B|A_i)P(A_i).$$

For continuous conditional probabilities, we replace the alternatives A_i by the "alternatives" (X=t), where t takes on all real values, and we replace the sum by an integral.

For any continuous random variable X and event A for which the continuous conditional probabilities P(A|X=t) exist,

$$P(A) = \int_{\infty}^{\infty} P(A|X=t) \ \mathrm{dens}(X=t)dt$$

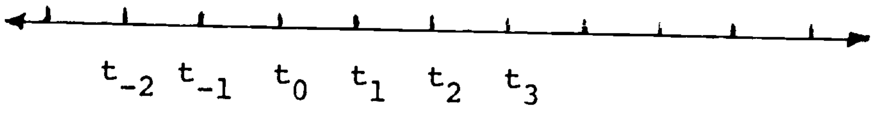
Continuous Law of Alternatives

We will give a rigorous proof of this law. The key fact we need is the Mean Value Theorem of Calculus. Recall what this says. If f is a continuous function on the interval [a,b], then for some point $\bar{x}$ between a and b,

$$f(\bar{x}) = \frac{1}{b-a} \int_a^b f(x)\,dx.$$

Proof of the Continuous Law of Alternatives

Let $\varepsilon>0$ be a small number. Divide up the real line into intervals of length ε by the points $t_n = n\varepsilon$.



Take B_i to be the event $(t_i<X\leq t_{i+1})$. Then the B_i form a set of alternatives. By the (ordinary) law of alternatives,

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

$$= \sum_i P(A|t_i<X\leq t_i+\varepsilon)P(t_i<X\leq t_i+\varepsilon)$$

By the mean value theorem applied to $f(t) = \text{dens}(X=t)$, there is some $\bar{t}_i$ in the interval $[t_i,t_{i+1}]$ such that

$$f(\bar{t}_i) = \frac{1}{\Delta t_i} \int_{t_i}^{t_{i+1}} f(t)dt = \frac{1}{\Delta t_i} P(t_i<X\leq t_i+\varepsilon)$$

or $P(t_i<X\leq t_i+\varepsilon) = f(\bar{t}_i)\Delta t_i$.

Therefore

$$P(A) = \sum_i P(A|t_i<X\leq t_i+\varepsilon)f(\bar{t}_i)\Delta t_i.$$

Now as $\varepsilon\to 0$ this last sum approaches

$$\int_{-\infty}^{\infty} P(A|X=t)f(t)dt,$$

by the definition of the integral. We have therefore proved the continuous law of alternatives.

Notice that we used the fact that the density of X, f(t) = dens(X=t), is continuous. Actually, in practice it will only be piecewise continuous. This technical detail will never bother us. The continuous law of alternatives holds in this case also.

6. Conditional Densities

In most computations concerning continuous random variables, the densities are much easier to handle. We can give density versions of conditional probability, continuous conditional probability and the continuous law of alternatives.

Let us begin with the simplest case: conditional density. Suppose we have an event B such that P(B)>0 and a random variable Y. The distribution of Y given that B has occurred is

$$F(s) = P(Y \leq s | B).$$

In general it is possible that a continuous random variable can fail to be continuous after conditioning, as we saw in the previous section. But if it is still continuous, we may speak of the conditional density of Y given B:

$$\mathrm{dens}(Y=s|B) = F'(s) = \frac{d}{dy} P(Y \leq s | B).$$

If the event B is of the form (X=t), then the conditional density can be defined by a limiting process just as we did in the last section. More precisely, the <u>continuous conditional density of</u> Y <u>given</u> X = t is

$$\text{dens}(Y=s|X=t) = \lim_{\varepsilon\to 0} \text{dens}(Y=s|t<X\leq t+\varepsilon) \quad ,$$

if this limit exists. If dens(X=t)≠0, then

$$\text{dens}(Y=s|X=t) = \frac{\text{dens}((Y=s)\cap(X=t))}{\text{dens}(X=t)} \quad ,$$

exactly as one would expect.

Consider again the example of section 5. The conditional density $\text{dens}(X_{(1)}=x|X_{(k)}=t)$, when $x<t$, is the same as the density $\text{dens}(X_{(1)}=x)$ but for the process of dropping k-1 points on [0,t], i.e.

$$\text{dens}(X_{(1)} = x|X_{(k)}=t) = \frac{(k-1)(t-x)^{k-2}}{t^{k-1}} \ .$$

We could also compute this as follows:

$$\text{dens}(X_{(1)}=x|X_{(k)}=t) = \frac{\text{dens}((X_{(1)}=x)\cap(X_{(k)}=t))}{\text{dens}(X_{(k)}=t)} \ .$$

Two two densities on the right were computed in section III.3:

$$\text{dens}(X_{(k)}=t) = n\binom{n-1}{k-1} \frac{t^{k-1}(a-t)^{n-k}}{a^n}$$

$$\text{dens}((X_{(1)}=x)\cap(X_{(k)}=t)) = \binom{n}{0,1,k-2,1,n-k} \frac{(t-x)^{k-2}(a-t)^{n-k}}{a^n}$$

Therefore,

$$\text{dens}(X_{(1)}=x|X_{(k)}=t)$$

$$= \frac{\binom{n}{0,1,k-2,1,n-k}(t-x)^{k-2}(a-t)^{n-k}}{n\binom{n-1}{k-1}t^{k-1}(a-t)^{n-k}}$$

$$= \frac{(k-1)(t-x)^{k-2}}{t^{k-1}}$$

The Continuous Bayes' Law

By using conditional densities one can formulate a continuous version of Bayes' law. Suppose we have two random variables X and Y. We call X the "cause" and Y the "effect". For example X may represent a parameter in an experiment, which we cannot measure directly, while Y is some directly measureable quantity. We want to determine the effect on the distribution of X given a particular observation of Y. As in the discrete Bayes' law, we assume that we know the a priori distribution of X, dens(X=x), as well as the conditional densities of Y given a value of X, dens(Y=y| X=x). By a calculation almost identical to the one for the discrete Bayes' law, we have this formula:

$$\text{dens}(X=x| Y=y) = \frac{\text{dens}(Y=y| X=x)\text{dens}(X=x)}{\int_{-\infty}^{\infty}\text{dens}(Y=y|X=t)\text{dens}(X=t)dt}$$

Continuous Bayes' Law

Continuous Law of Successive Conditioning

In a similar manner as that above, one can state a continuous analog of the law of successive conditioning. We leave the details as an exercise.

If $X_1, X_2, \dots, X_n$ are a sequence of continuous random variables, then their joint density is given by:

$$\text{dens}(X_1=t_1, X_2=t_2, \dots, X_n=t_n) =$$
$$= \text{dens}(X_1=t_1)\text{dens}(X_2=t_2|X_1=t_1)\cdots$$
$$\cdots\text{dens}(X_n=t_n|X_1=t_1, \dots, X_{n-1}=t_{n-1})$$

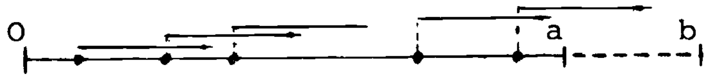
Continuous Law of Successive Conditioning

7. Gaps in the Uniform Process

As an application and illustration of the conditioning techniques just introduced, we give a detailed and rigorous treatment of the gaps L_i in the Uniform process. We begin with a problem posed in the introduction. Namely, if we drop a set of needles, each of length h, on a stick of length b, what is the probability that none of the needles overlap?

Needles on a Stick

We first restate the problem in terms of the Uniform process. The position of a given needle is completely determined by its left endpoint. The process of dropping n needles of length h on a stick of length b is then the same as dropping n points on the interval [0,b-h]. Write a=b-h.



5 needles of length h on a stick of length b

= 5 points on an interval of length a = b-h

Now two needles are non-overlapping if and only if their left endpoints are at least distance h apart. Let $A_{a,n}$ be the event "n needles on a stick of length b = a+h do not overlap". Then

$$P(A_{a,n}) = P((L_2 \geq h) \cap (L_3 \geq h) \cap \ldots \cap (L_n \geq h))$$

$$= P\left(\left(\min_{2 \leq i \leq n} L_i\right) \geq h\right)$$

We will first compute the probability of a slightly different event. Let $B_{a,n}$ be the event "n points dropped on [0,a] are all at least distance h from each other <u>and</u> from the right endpoint." This is exactly the same as $A_{a,n}$ but we have added the condition that the last gap, L_{n+1}, also be larger than h, i.e.

$$A_{a,n} = (L_2 \geq h) \cap (L_3 \geq h) \cap \ldots \cap (L_n \geq h)$$

$$B_{a,n} = (L_2 \geq h) \cap (L_3 \geq h) \cap \ldots \cap (L_n \geq h) \cap (L_{n+1} \geq h).$$

To compute $P(B_{a,n})$ we condition on the position of the largest point:

$$P(B_{a,n}) = \int_{-\infty}^{\infty} P(B_{a,n} | X_{(n)} = t) \,\text{dens}(X_{(n)} = t)\, dt$$

$$= \int_{(n-1)h}^{a-h} P(B_{a,n} | X_{(n)} = t) \frac{n}{a^n} t^{n-1} \, dt$$

Here the limits of integration stem from the fact that $B_{a,n}$ can only occur if the largest point falls so that the rightmost gap, L_{n+1}, is larger than h and so that there is enough "room" for n-1 gaps of size h to the left of $X_{(n)}$. Now $P(B_{a,n}|X_{(n)}=t)$ is the same as the probability of dropping n-1 points on [0,t] so that all gaps are at least h and also so that the largest point, $X_{(n-1)}$, falls at least distance h from t. Thus

$$P(B_{a,n}|X_{(n)}=t) = P(B_{t,n-1}).$$

We may therefore use mathematical induction. For if we write $p_n(a) = P(B_{a,n})$, then

$$p_n(a) = \int_{(n-1)h}^{a-h} p_{n-1}(t)\frac{n}{a^n}\, t^{n-1}dt.$$

Consider $p_1(a)$. This is the probability that a point dropped on [0,a] falls farther than distance h from a. So $p_1(a) = \frac{a-h}{a}$. More generally the above inductive formula can be used to deduce that $p_n(a) = \left(\frac{a-nh}{a}\right)^n$.

To compute $P(A_{a,n})$ from what we know about $P(B_{a,n})$, we condition on the largest point $X_{(n)}$. It is easy to see that $P(A_{a,n}|X_{(n)}=t) = P(B_{t,n-1})$. Therefore, as above,

$$P(A_{a,n}) = \int_{-\infty}^{\infty} P(A_{a,n}|X_{(n)}=t)\,\mathrm{dens}(X_{(n)}=t)\,dt$$

$$= \int_{(n-1)h}^{a} P(B_{t,n-1})\frac{n}{a^n}t^{n-1}\,dt$$

$$= \frac{n}{a^n}\int_{(n-1)h}^{a} \left(\frac{t-(n-1)h}{t}\right)^{n-1} t^{n-1}\,dt$$

$$= \frac{n}{a^n}\left[\frac{(t-(n-1)h)^n}{n}\right]_{(n-1)h}^{a}$$

$$= \frac{(a-(n-1)h)^n}{a^n}$$

$$= \left(\frac{b-nh}{b-h}\right)^n \qquad \text{(provided } b\geq nh\text{)}$$

<u>Exchangeability</u> <u>of</u> <u>the</u> <u>gaps</u>

Recall that we stated in section II.3 that the gaps L_i in the Uniform process are equidistributed but that we gave only an intuitive justification. We can now give a rigorous proof using conditional density.

Consider the first two gaps. The density of L_1 is $\text{dens}(L_1=t_1) = \frac{n(a-t_1)^{n-1}}{a^n}$. Therefore, by the law of alternatives,

$$\text{dens}(L_2=t_2) = \int_{-\infty}^{\infty} \text{dens}(L_2=t_2|L_1=t_1)\text{dens}(L_1=t_1)dt_1.$$

Now the conditional density $\text{dens}(L_2=t_2|L_1=t_1)$ is the same as that of the <u>first</u> gap in the Uniform process of dropping <u>n-1</u> points on the interval $[t_1,a]$. Therefore

$$\text{dens}(L_2=t_2|L_1=t_1) = \begin{cases} \frac{(n-1)(a-t_1-t_2)^{n-2}}{(a-t_1)^{n-1}} & \text{if } 0\le t_2\le a-t_1 \\ 0 & \text{otherwise} \end{cases}$$

Hence:

$$\text{dens}(L_2=t_2) = \int_0^{a-t_2} \frac{(n-1)(a-t_1-t_2)^{n-2}}{(a-t_1)^{n-1}} \cdot \frac{n(a-t_1)^{n-1}}{a^n} dt_1$$

$$= \int_0^{a-t_2} \frac{(n-1)(n)}{a^n} (a-t_1-t_2)^{n-2}dt_1$$

$$= \frac{(n-1)(n)}{a^n} \left[- \frac{(a-t_1-t_2)^{n-1}}{n-1} \right]_0^{a-t_2}$$

$$= \frac{n}{a^n} (a-t_2)^{n-1}$$

Therefore L_1 and L_2 are equidistributed.

We end with an example of the use of the law of successive conditioning. We compute the joint density of the gaps $L_1,\dots,L_{n+1}$:

$$\mathrm{dens}(L_1=t_1,\ L_2=t_2,\dots,\ L_{n+1}=t_{n+1}) =$$

$$= \mathrm{dens}(L_1=t_1)\cdot\mathrm{dens}(L_2=t_2 \mid L_1=t_1)\cdot\mathrm{dens}(L_3=t_3 \mid L_1=t_1,L_2=t_2)\cdots$$

The conditional density

$$\mathrm{dens}(L_j=t_j \mid L_1=t_1,L_2=t_2,\dots,L_{j-1}=t_{j-1})$$

is the same as the density of the first gap in the Uniform process of dropping n-j+1 points on $[t_1+\dots+t_{j-1},a]$:

$$\frac{(n-j+1)(a-t_1-t_2-\dots-t_j)^{n-j}}{(a-t_1-t_2-\dots-t_{j-1})^{n-j+1}}$$

Therefore

$$\mathrm{dens}(L_1=t_1,\ L_2=t_2,\dots,\ L_{n+1}=t_{n+1}) = \begin{cases} \dfrac{n!}{a^n} & \text{if } t_1+\dots+t_{n+1}=a \\ 0 & \text{otherwise} \end{cases}$$

by cancellation.

The joint density of any collection of gaps can be computed from the above formula by taking marginals with respect to all the other gaps. As a result we see that all the gaps are equidistributed. Even more is true: the gaps are exchangeable! At first this does not seem correct, but it is possible to see it intuitively if we return to the "points on a circle" interpretation of the Uniform process as in section III.3.

An immediate, and by no means obvious, consequence of the exchangeability of the gaps is that the covariance of any pair of them is the same as that of the first two. This implies paradoxically that the correlation between the first two gaps is the same as that between L_1 and L_i for any i!

Table of Conditioning Laws

	Probabilities	Densities
Conditioning	$P(A \mid B) = \frac{P(A\cap B)}{P(B)}$	$\text{dens}(X=t \mid B) = \frac{d}{dt} P(X\leq t \mid B)$
Continuous Conditioning	$P(A \mid Y=s) =$ $= \lim_{\varepsilon\to 0} P(A \mid s<Y\leq 1+\varepsilon)$	$\text{dens}(X=t \mid Y=s) = \frac{\text{dens}((X=t)\cap(Y=s))}{\text{dens}(Y=s)}$ $= \lim_{\varepsilon\to 0} \frac{d}{dt} P(X\leq t \mid s<Y\leq s+\varepsilon)$

Types of Conditioning

	Probabilities	Densities
Conditioning	$P(B) = \sum_i P(B \mid A_i) P(A_i)$	$\text{dens}(Y{=}s) = \sum_i \text{dens}(Y{=}s \mid A_i) P(A_i)$
Continuous Conditioning	$P(B) = \int_{-\infty}^{\infty} P(B \mid X{=}t)\,\text{dens}(X{=}t)dt$	$\text{dens}(Y{=}s) =$ $\int_{-\infty}^{\infty} \text{dens}(Y{=}s \mid X{=}t)\,\text{dens}(X{=}t)\,dt$

Law of Alternatives

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_j P(B \mid A_j)P(A_j)}$$

$$\text{dens}(X{=}x \mid Y{=}y) = \frac{\text{dens}(Y{=}y \mid X{=}x)\text{dens}(X{=}x)}{\int_{-\infty}^{\infty} \text{dens}(Y{=}y \mid X{=}t)\text{dens}(X{=}t)dt}$$

Bayes' Law

$$P(B_1 \cap B_2 \cap \ldots \cap B_n) = P(B_1)P(B_2 \mid B_1)P(B_3 \mid B_1 \cap B_2)\ldots P(B_n \mid B_1 \cap B_2 \cap \ldots \cap B_{n-1})$$

$$\text{dens}(X_1{=}t_1, X_2{=}t_2, \ldots, X_n{=}t_n) =$$

$$= \text{dens}(X_1{=}t_1)\text{dens}(X_2{=}t_2 \mid X_1{=}t_1)\ldots \text{dens}(X_n{=}t_n \mid X_1{=}t_1, \ldots, X_{n-1}{=}t_{n-1})$$

Law of Successive Conditioning

8. The Algebra of Probability Distributions

Very early in our study of random variables we noted that we can perform algebraic operations on them to get new random variables. We did not, however, make any systematic study of what effect an algebraic operation has on the distribution of the random variables involved. For example, if X is a continuous random variable with density f(x), what is the density of 2X? The answer is most assuredly not 2f(x). In fact, the correct answer is $\frac{1}{2}f(\frac{x}{2})$. This illustrates a basic fact about algebraic operations on random variables: the effect of a simple operation on a random variable is seldom reflected in a simple way on its density. In this section we consider two kinds of operations on random variables: "change of variables" on a single random variable and the sum of two independent random variables.

Change of Variables

Let X be a continuous random variable, whose density is f(x). Let g(x) be an increasing function. We wish to determine the distribution of the random variable g(X). To do so we must consider the distribution function of X, not just its density. Accordingly, let F(x) be $P(X \leq x)$, so that $f(x) = F'(x)$. The distribution function of g(X) is given by $P(g(X) \leq x)$. Now we assumed that g(x) was an increasing function, so it has an inverse function G(y) which is also increasing. Therefore,

$$P(g(X) \leq x) = P(G(g(X)) \leq G(x)) = P(X \leq G(x)) = F(G(x)).$$

To get the density of $g(X)$ we differentiate this, using the chain rule:

$$\text{dens}(g(X)=x) = F'(G(x))G'(x) = f(G(x))G'(x).$$

By the inverse function principle of Calculus, $G'(x) = 1/g'(G(x))$. Therefore, we have shown:

$$\boxed{\begin{array}{c} \text{dens}(g(X)=x) = \dfrac{f(G(x))}{g'(G(x))} \\[1em] \text{Change of Variables Formula} \end{array}}$$

An immediate consequence of this formula is that for any continuous random variable X with distribution function F(x), F(X) is uniformly distributed on $[0,1]$. Thus every continuous random variable is, by a change of variables, expressible in terms of any other. This fact can be used in computer simulations of stochastic processes. Most computer systems provide a pseudo-random number generator which produces, with each call, an independent, uniformly distributed pseudo-random number from $[0,1]$. Call this number RND. If we wish to simulate a random variable X whose distribution function is F(x), we just use G(RND), where G(y) is the inverse function of F(x).

The change of variables formula we have given applies only to an increasing function g(x). For a decreasing function, the only change is that the sign of the right-hand side must be reversed. For more complicated functions g(x), one must partition the domain of g(x) into intervals on which it is increasing or decreasing and apply the change of variables formula to each such interval. The results must then be combined to get the density of $g(X)$. Needless to say this can get quite intricate.

Sums of Independent Random Variables

Suppose that X and Y are two random variables. Their sum is again a random variable, X+Y. For example, in the uniform process, $X_{(2)} = L_1+L_2$. Now if we know the distributions of X and Y, can we compute the distribution of the sum X+Y? In general, the answer is no, for we need the joint distribution in order to compute the distribution of the sum. In the above example, we cannot compute the distribution of $X_{(2)}$ from the distributions of L_1 and L_2 alone, we must know also the joint distribution of L_1 and L_2.

On the other hand, if X and Y are independent, we can compute their joint distribution from their individual distributions. As a result we expect that the distribution of X+Y has some reasonable expression in terms of the distributions of X and Y. Suppose for the moment that X and Y are independent, integer random variables with distributions

$$P(X=n) = p_n$$

$$P(Y=n) = q_n$$

Then by the law of alternatives,

$$P(X+Y=n) = \sum_k P(X+Y=n|Y=k)P(Y=k)$$

$$= \sum_k P(X=n-k|Y=k)P(Y=k).$$

Since X and Y are independent, $P(X=n-k|Y=k) = P(X=n-k)$. Therefore,

$$P(X+Y=n) = \sum_k P(X=n-k)P(Y=k)$$

$$= \sum_k p_{n-k}\, q_k$$

The distribution $r_n = \sum_k p_{n-k}q_k$ is called the (discrete) convolution of the distributions p_n and q_n.

In the case of integer random variables, we can see clearly what the convolution means: $P(X+Y=n)$, is the sum of all possible "ways" that X+Y can equal n: X=k and Y=n-k, for all possible k. In the continuous case, the sum is replaced by an integral, but the idea is the same.

Suppose that X and Y are continuous random variables having densities

$$\text{dens}(X=x) = f(x)$$

$$\text{dens}(Y=x) = g(x).$$

$$\text{dens}(X+Y=t) = \int_{-\infty}^{\infty} \text{dens}(X+Y=t|X=s)\,\text{dens}(X=s)\,ds$$

$$= \int_{-\infty}^{\infty} \text{dens}(Y=t-s|X=s)\,f(s)\,ds$$

$$= \int_{-\infty}^{\infty} \text{dens}(Y=t-s) f(s) ds$$

$$= \int_{-\infty}^{\infty} g(t-s) f(s) ds.$$

The function

$$h(t) = \int_{-\infty}^{\infty} g(t-s) f(s) ds$$

is called the convolution of f and g, which we shall write

$$h = f*g .$$

We have just proved that: a sum of independent continuous random variables corresponds to convolution of the densities.

The convolution of two functions is an important operation which appears in numerous contexts, for example dynamical systems in engineering and optics in physics, to name just two. Its appearance in probability theory is perhaps the most easily understood context in which the convolution arises. Actually there are many operations similar to the one above that go by the name "convolution." For example if we consider the special case of two continuous, positive, independent random variables X and Y, the density of their sum is

$$h(t) = \int_{0}^{t} g(t-s) f(s) ds ,$$

because $f(s) = 0$ if $s<0$ and $g(t-s) = 0$ if $s>t$. This is the form of the definition one sees most commonly.

Although it is not obvious from the definition, convolution is a commutative, associative operation. That is, for densities f,g and h:

$$f*g = g*f$$

$$(f*g)*h = f*(g*h).$$

These follow from the fact that addition of random variables is commutative and associative, respectively.

As an example of a convolution, we show a result which is implicit in many of the calculations in chapter IV: the sum of normally distributed random variables is again normal. We will just take the case of two standard normal random variables, and leave the general case as an exercise. By definition, the convolution of the standard normal density with itself is given by:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)\cdot\frac{1}{\sqrt{2\pi}}\exp(-(y-x)^2/2)\ dx$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left[\frac{-x^2-y^2+2xy-x^2}{2}\right] dx$$

$$= \frac{1}{2\pi} \exp(-y^2/2) \int_{-\infty}^{\infty} \exp\left[\frac{-2x^2+2xy}{2}\right] dx$$

$$= \frac{1}{2\pi} \exp(-y^2/2) \int_{-\infty}^{\infty} \exp\left[-x(x-y)\right] dx$$

Using the change of variables t=x-y/2, this becomes:

$$\frac{1}{2\pi}\exp(-y^2/2)\int_{-\infty}^{\infty}\exp\left[-(u+y/2)(u-y/2)\right]du$$

$$= \frac{1}{2\pi}\exp(-y^2/2)\int_{-\infty}^{\infty}\exp(-u^2+y^2/4)\,du$$

$$= \frac{1}{2\pi}\exp(-y^2/2+y^2/4)\int_{-\infty}^{\infty}\exp(-u^2)\,du$$

$$= \frac{1}{2\pi}\exp(-y^2/4)\sqrt{\pi}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\exp(-y^2/(2\sigma^2)), \quad \text{where } \sigma = \sqrt{2}.$$

9.* Geometric Probability

10. Exercises for

Chapter V Conditional Probability

Discrete Conditional Probability

1. A game is played with six double-sided cards. One card has "1" on one side and "2" on the other. Two cards have "2" and "3" on the two sides. And the last three have "3" and "4" on them. The six cards are shuffled by one person. A random card is then drawn and held in a random orientation between two other persons, each of whom sees only one side of the card. The winner is the one seeing the smaller number. Suppose that the first person chooses the "2/3" card. Compute the probabilities each of the two persons thinks he/she has for winning.

2. A person is given an urn and is told it contains 4 balls: 2 red and 2 black. He draws two of the balls at random without replacing them, and both turn out to be red. He puts these aside. What is the probability that the next ball drawn is black? Another person in the room has been blindfolded during all of the preceding. After taking off her blindfold, she takes a ball out of the urn at random. She knows which balls were originally in the urn and that two have been drawn so far but does not know their color. What does she think the probability of drawing a black ball is? How could the fact that she was blindfolded have any effect on the probability of the next drawing of a ball? Explain.

3. Place k balls into n boxes at random. If the first box is empty, what is the probability that the second is also?

4. During a poker game a kibitzer manages to get a brief glimpse of one of the hands (and no other hands). In this glimpse he sees only that one card in the hand is an ace. He did not notice which ace it was. What is the probability that the hand has at least two aces? If the kibitzer noticed that one card is a black ace, what is the probability that the hand has at least two aces? Finally suppose the kibitzer saw that the hand had the Ace of Spades. Discuss whether such glimpses are really possible. The "moral" of this example is that (conditional) probabilities of events change considerably when one learns kinds of information that have no obvious relevance.

5. Three prisoners are informed by their jailer that one of them has been chosen at random to be executed and that the other two are to be freed. They are told they will learn their fate in one week's time. Prisoner A asks the jailer to tell him privately the name of a fellow prisoner who will be set free, claiming that there would be no harm in divulging this information, since he already knows that at least one will go free, and he cannot inform the prisoner in question about his good fortune. The jailer refuses to tell prisoner A, pointing out that if A knew the name of one of his fellows to be set free, then his own probability of being executed would rise from 1/3 to 1/2, since he would then be one of two prisoners; and this would be cruel. What do you think of the jailer's reasoning? Be precise.

6. You are playing bridge. Assume that the deck is thoroughly shuffled. If you receive 4 hearts, how many did your partner receive? Generalize to the case of receiving N hearts.

7. (Neyman-Pearson errors) A commuter has the choice of taking the train to work or of driving to work. If she takes the train she will get to work on time about one time in four. If she takes her car, she is almost certain of getting to work on time, but at considerable inconvenience. Although she calls the transit company every morning, their information is wrong a third of the time. So she adopts the following strategy: if the transit company says the train will be late, she always takes her car, and if not she takes her car a third of the time at random. Compute how probable it is that she will be late. What is the probability that she takes her car even though she would have been on time if she had taken the train? This kind of "mistake" is known as a Neyman-Pearson Error of Type I. She makes an Error of Type II if the train is late when she takes it. Compute the probability that she makes an error of type II. Note that the probability of either kind of error is a conditional probability. Furthermore in order to make the above computations, one must assume a number of independence properties of the various events. State explicitly any assumptions you must make in this problem. The probability of an error of type I is called the significance level of the decision, and 1 minus the probability of an error of type II is called the power of the decision. One should re-examine statistical hypothesis testing using this terminology. When several tests are available one clearly would like the one with the largest power for a given significance level. Unfortunately, however, in practice one rarely knows precisely what model will be implied by the rejection of the hypothesis being tested, so computing the power of a test is not as easy as it seems.

8 . Suppose that the commuter in exercise 7 scores the two inconveniences of being late and of driving the car at 1 and 2 respectively. What is her optimal strategy? Do the same with 1 and 2 interchanged.

9. An event A of positive probability is said to be favorable to an event B if

$$P(B|A) \geq P(B),$$

in other words, if we know that A has occurred then the probability that B has occurred also is the same as it was or is greater. Note that if A is independent of B, then it is favorable to B. Suppose we have a family having two children. Let A be the event "the first child is a girl," let B be the event "the second child is a girl," and let C be the event "the two children have different gender." Show that A and B are both favorable to C but that $A \cap B$ is not. Similarly show that C is favorable to both A and B but not to $A \cap B$. Give examples to show that an event can be favorable to two others without favoring their union and conversely that two events can favor a third without their union doing so.

10. (Simpson's paradox) A new treatment for a disease has just become available but is still experimental and is very expensive. In a teaching hospital with a large budget a random sample of 100 patients with the disease are randomly broken into two groups, one having 90 patients the other 10. The larger group is treated. 30 of these show clear improvement and only one of the members of the other group does. In a city hospital with a smaller budget a

similar test is made but now the smaller group gets the treatment. In this group 9 show improvement while in the untreated group half improve and half do not. In either case the treatment seems to be effective. However, if we view this as one sample of 200, 100 of which are treated and the other 100 are not, then a different picture emerges. Of the treated patients 39 improve and of the untreated patients some 46 improve. This seems to suggest that the treatment actually decreases ones chance for improvement. Explain the apparent paradox here.

Bayes' Law

11. There are three children in a family. A friend is told that at least two of them are boys. What is the probability that all three are boys? The friend is then told that the two are the oldest two children. Now what is the probability that all three are boys? Use Bayes' Law to explain this. Assume throughout that boys are as likely as girls and that each child is independently either a boy or a girl.

12. A student is about to take a quiz. If he studies, he will pass with probability .99, but if he goes to the dorm party his chances of passing decline to 1/2. The next day he passes the exam. Did he go to the party?

13.* Use Bayes' law to compute the probability in exercise 7 that the commuter took the train given that she was on time.

14. Three boxes each contain two coins. One has two silver coins, one has two gold coins, and one has one of each. A box is chosen completely at random and a coin is chosen at random from that box. It is gold. Is the other coin in the box gold also?

15. The manufacturer of screws in exercise IV.15 is producing good screws 99% of the time but now the machine that detects the flawed screws is itself out of adjustment, producing an incorrect decision 10% of the time. What is the probability that a discarded screw is really flawed?

16. A lie detector test is known to be 80% reliable when the person is quilty and 95% reliable when the person is innocent. If a suspect was chosen from a group of suspects of which only 1% have ever committed a crime, and the test indicates that he is quilty, what is the probability that he is innocent?

17. In the optimal choice problem (exercise III.64) the correct strategy is to make no decision for a certain length of time (say k days) and then to choose the best candidate of all those seen up to that point. If the monarch chooses the j^{th} candidate what is the probability that she is the best candidate?

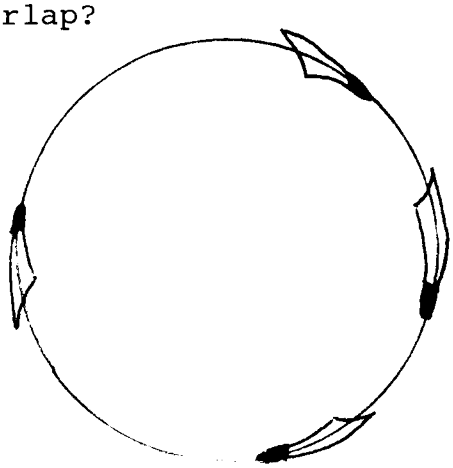
18*. In exercise 17 above compute the probability that the monarch will choose the j^{th} candidate using the above strategy. Use this to find the probability that this strategy succeeds. For which k will this be maximized? Generalize to N candidates.

19.* Use Bayes' law to compute the probability of each of the kinds of hands in exercise II.21 given that one has at least one pair.

Continuous Conditional Probability

20. A target is a disk of radius 1m. A bullet is fired at the disk and hits it. Assume the bullet's mark has a uniform distribution, i.e., the probability that it hits a region A is proportional to the area of A. How far from the center does the bullet hit?

21. A scimitar is a sward shaped like a circular arc (at least for this problem). Suppose that during a Turkish festival n Turks throw their scimitars independently and at random along a circle of circumference a. Suppose that each scimitar has arc length h along this circle. What is the probability that none of the scimitars overlap?



Circle of circumference a with
4 nonoverlapping scimitars.

22. In the Uniform process of sampling n > 2 points from [0,a], what is the probability that the first three gaps are all less than b ?

23. For $1 \leq i < j < k \leq n$, compute the joint density of $X_{(i)}$ and $X_{(k)}$ given that $X_{(j)} = t$.

24. Let $L_1, L_2, \cdots, L_{n+1}$ be the gaps in the Uniform process of sampling n points from $[0,a]$. Find the distributions of the order statistics of the gaps, i.e., put the gaps in order, getting the random variables $L_{(1)}, L_{(2)}, \cdots, L_{(n+1)}$. Then compute their expectations. Compare this with the scimitar problem (exercise 21) above and with the broken DNA problem (exercise III.53).

25. Compute the distribution of the median gap in the Uniform process of sampling n points from $[0,a]$.

26. In the Uniform process of sampling n points from $[0,a]$, what is the probability that the largest gap is at least twice as large as the smallest gap?

27. Given positive numbers $t_1, t_2, \ldots, t_{n+1}$, what is the probability that in the uniform process, for all $i = 1, 2, \ldots, n+1$, the i^{th} gap is greater than t_i.

28*. Give a rigorous statement and proof of the identity

$$\text{dens}(X = t \mid Y = s) = \frac{\text{dens}(X = t, Y = s)}{\text{dens}(Y = s)}$$

29*. Give a rigorous $(\varepsilon - \delta)$ proof of the continuous law of alternatives.

30*. Define a <u>cluster of size</u> k <u>and width</u> ε to be a sequence of k points contained in an interval of length ε. In the Uniform process how many clusters of size k and width ε are there?

Exchangeability

31. Drop r red points and b black points $(r+b=n)$ at random uniformly on $[0,a]$. What is the probability that a run of h red points precedes a run of t black points?

32. Compute the probability that at least one of the four players in a bridge game is dealt a yarborough. Note that the four hands are not independent but are exchangeable. Compare this answer with what you would get if you dealt the four hands independently from four different decks. Use the result of exercise II.12.

33.* (Discrete Needles on a Stick Problem). Choose k numbers from the set $\{1,2\ldots,n\}$ at random. What is the probability that no two are closer than a units apart? Note that the answer depends on whether one uses Fermi-Dirac, Bose-Einstein or Maxwell-Boltzmann statistics.

34.* (Discrete Scimitars on a Circle Problem). Choose k numbers from the set of integers modulo n. What is the probability that no two have a difference congruent modulo n to one of the integers in the set $\{-a+1,\ldots,-2,-1,0,1,\ldots,a-1\}$? As in exercise 33 above, the depends on which kind of statistics we use.

Change of Variables

35. Find the distributions of the following random variables in terms of that of the random variable X:

(a) $Y = X + c$, where c is a constant,

(b) $Y = aX + b$, where a and b are constants,

(c) $Y = |X|$

35. (Continued)

(d) $Y = \sqrt{X}$, where X is a positive random variable ,

(e) $Y = \ln(X)$, where X is a positive random variable and $\ln$ denotes the natural logarithm.

36. A point is dropped at random (uniformly) on a square of side a. What is the distance of this point from the center of the square?

37. Let S be the speed of a molecule in a uniform gas at equilibrium. Then S is a positive random variable whose density is given by $\text{dens}(S = s) = 4\sqrt{b^3/\pi}\; s^2 e^{-bs^2}$ for $s > 0$, where b is a constant which depends on the absolute temperature and mass of the molecule. Find the density of the kinetic energy E of the molecule, $E = \frac{1}{2} m S^2$.

38. Suppose that a long DNA molecule of length a is broken at random into two pieces. Compute the distribution of the ratio of the length of the longer piece by that of the shorter piece. Compute the ratio of the expected sizes of the longer and shorter pieces and the expected ratio of the longer and shorter pieces. [Answers: 3 and ∞]. Do the same for a molecule broken into 3 pieces.

39. (Student's t-distribution) The key fact behind much of modern statistical theory is the Central Limit Theorem: the standardization of a sum of independent, equidistributed random variables is normally distributed in the limit as the sample size gets large. Now we remarked in chapter IV that if we do not know the variance of the random variables, then we can approximate

it using the sample variance. Unfortunately, if we only have a samll sample, the fact that the sample variance is being used instead of the actual variance can result in the standardized sum having a distribution considerably different from a normal distribution even if the original sequence of random variables were all normally distributed. This fact was first pointed out by William Gosset, who wrote under the pseudonym of "Student." We will now consider his computation.

Let $X_1, \cdots, X_n$ be a sequence of independent, normally distributed R.V.'s with mean m and variance σ^2. The sample mean is $\bar{m} = \frac{1}{n}(X_1 + \cdots + X_n)$ and the sample variance is $\bar{\sigma}^2 = \frac{1}{n-1}((X_1 - \bar{m})^2 + \cdots + (X_n - \bar{m})^2)$. We wish to compute the distribution of the random variable $t = (\bar{m} - m)\frac{\sqrt{n}}{\bar{\sigma}}$. Note that this R. V. is not defined for $n = 1$. For the purposes of computing the distribution of t we may assume that $X_1, \cdots, X_n$ have distribution $N(0,1)$. Then for $n = 2$, t is the random variable $\frac{X_1 + X_2}{|X_1 - X_2|}$. It is easy to check that $X_1 + X_2$ and $X_1 - X_2$ both have distribution $N(0,2)$ and are independent. More generally show that t has the same distribution as the ratio

$$\frac{X}{\sqrt{Y_1^2 + \cdots + Y_{n-1}^2}} ,$$

where $X, Y_1, \cdots, Y_{n-1}$ are independent and have the standard normal distribution. The distribution of t is called the Student's t-distribution with $n-1$ degrees of freedom.

When $n = 2$, the distribution of t is the same (up to a scale change) as the gangster distribution in exercise III.xx. See exercise xx below. Compute the Student's t-distribution explicitly for the case of 1 degree of freedom.

40.* Let X and Y be independent, uniformly distributed random variables on $[0,1]$. Prove that $\cos(2\pi X)\sqrt{-2\ln(Y)}$ has distribution $N(0,1)$. This fact is useful for generating a sequence of independent, normally distributed pseudo-random numbers by computer, since most computers have a pseudo-random number generator that produces a sequence of independent , uniformly distributed pseudo-random numbers from $[0,1]$.

41.* Let $X_1, \cdots, X_n$ be a sequence of independent, standard normal random variables. Compute the distributions of the order statistics $X_{(1)}, \cdots, X_{(n)}$ of these random variables. Write a computer program that uses a numerical integration to find an approximation for $E(X_{(j)})$ accurate to 3 decimal places. Then make a table of $E(X_{(1)})$ for n between 1 and 20.

42.* In a college cafeteria ice cream is available for the evening meal in servings that vary in weight according to a normal distribution with a standard deviation of 100 gm. The cafeteria workers maintain about 15 servings for students to choose from. Every day student A chooses the smallest serving available while student D chooses the largest. Over the school year (200 meals), how much more ice cream does student D eat than student A?

43.* Drop n points at random independently and uniformly on a square of a side a. How close is the point closest to the center of the square?

44.** Drop $n+1$ points at random independently and uniformly on a square of side a. What is the distance to the nearest neighbor of the first point? (This is the pennies-on-a-carpet problem mentioned in the introduction and is currently an unsolved problem.)

Convolutions of Random Variables

45. (Rayleigh distribution) Let X and Y be independent random variables having distribution $N(0,\sigma^2)$. Find the distribution of $\sqrt{X^2 + Y^2}$. We can interpret this as the distribution of the deviation of an object from a target point when the object is dropped onto the target from above. X and Y are the deviations in the x and y directions with respect to rectangular coordinate system whose origin is at the target point.

46. For the situation described above, consider a circle and a square of the same area, both centered at the target point. Which is more likely to contain the point where the object lands? Hint: use probabilistic reasoning.

47. (χ^2 - distribution)

Return to exercise IV.19. The FDA should be just as concerned with variance as with the mean quantity of impurity. For example, if a company produces pills with an average of 4 ppm impurities but a variance of 4 $(\text{ppm})^2$, 31% of the pills it is producing are below standard. One can test a sample variance just as one can test a sample mean. The distribution of the sample variance of a random sample of size n from a normally distributed population is called the <u>chi-square distribution with</u> $n-1$ <u>degrees of freedom</u>. If the mean is known and not simply computed from the data of the random sample, then the distribution is the chi-square distribution with n degrees of freedom. See exercise 6. The chi-square distribution can be computed as follows. First compute the distribution of X^2, when X is $N(0,\sigma^2)$. This is the chi-square with one degree of freedom and mean σ^2. For the general case let $X_1, \cdots, X_n$ be independent random variables each distributed as $N(0,\sigma^2)$. The sum $\frac{1}{n}(X_1^2 + \cdots + X_n^2)$ has the distribution of the chi-square with mean σ^2. Compute the variance of the chi-square distribution. [Answer: $3\sigma^4$]. Now for large samples, we can use the Central Limit Theorem to conclude that the chi-square is approximately normally distributed. What is its variance? [Answer: $\frac{3\sigma^4}{n}$] Now suppose that the FDA determines that 9 ppm impurity is possibly hazardous. It would seem reasonable to require that no more than 1 pill per thousand can have this much impurity. Does the drug company examined in exercise IV.19 conform to this requirement? Use 95% one-sided confidence intervals both for the mean and for the variance. Note that the number of degrees of freedom in our sample is 49 not 50. [Answer: No]

48. Return to exercise IV.39. Suppose that the resisters are in parallel rather than series. Using a suitable normal approximation find a 95% confidence interval for the resistance of this circuit.

49.* In exercise III.36, the gangster sprays the wall with a machine gun, shooting n bullets independently, each in a random direction toward the wall. What is the distribution of the sum of the positions of the n bullet holes? Assume that the median (III.37) is the zero point and that distances are measured in meters. What is the distribution of the average position of the n bullet holes? Does this result explain what you observed in exercise IV.58?

50.* Compute the density of the sum of n independent, uniformly distributed random variables on [0,a].

51.* Give a rigorous proof of the convolution theorem.

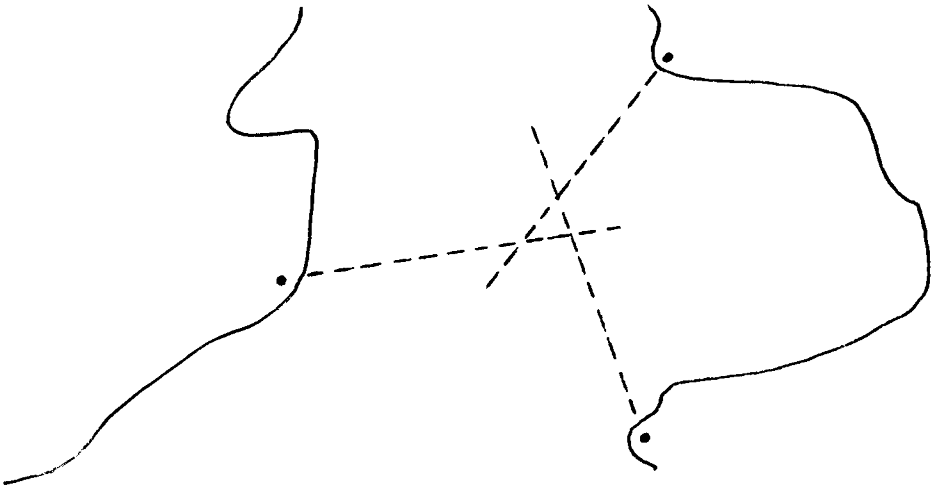
Geometric Probability

52. In a circus carnival game, a player tosses a quarter onto the surface of a table ruled in a checkerboard pattern of two-inch squares, which is further subdivided into one-inch squares by lines of another color. A quarter is $\frac{15''}{16}$ in diameter. If it falls entirely inside one of the two-inch squares, the player receives 50¢ (his original quarter plus another). If it falls inside one of the one-inch squares, he receives a prize worth about twenty dollars. Otherwise the player receives nothing. What is the probability of winning each prize, and what is the

average return on ones investment for one toss of a quarter?

53. Choose four points at random on a circle. Call them X_1, X_2, X_3 and X_4. What is the probability that the chords $\overline{X_1X_2}$ and $\overline{X_3X_4}$ intersect? Hint: use a symmetry argument.

54. A captain of a ship can determine its position by using radio bearings from transmitters on shore. These only give a direction so it is necessary to use at least two such bearings to determine the position. However, because of errors of measurement, one usually takes three bearings and the position lines are then plotted on a map as in the example below. The ship is assumed to lie inside the triangle formed by the three lines. All we know about the errors of measurement in the three bearings is that they are independent and symmetric about the true bearing. What is the probability that the ship actually lies inside the triangle? [Answer: 1/4] .



55. Let $\vec{X}$ be a randomly oriented unit vector in 3-space. Show that the length L of the projection of $\vec{X}$ on the x-axis (i.e., the x-component of $\vec{X}$) is uniformly distributed on [0,1] and that $E(L) = 1/2$.

56. (Feller) What is the length of a random segment intersecting a unit sphere? More precisely, let P be a point on the sphere, and let L be a line through P in a random direction. Let S be the length of the intersection of L with the sphere. What is the distribution of S? [Answer: uniformly distributed on [0,2]].

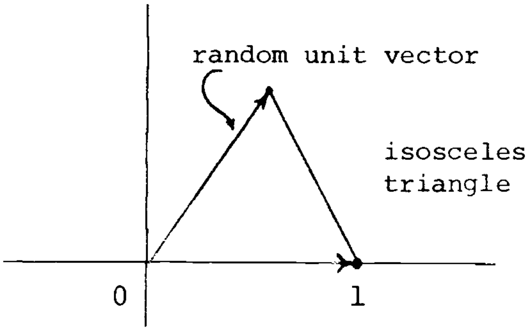
57. Let $\vec{X}$ be as in exercise 55, and let U be the length of the projection of $\vec{X}$ on the (x,y)-plane. Show that U has probability density $f(t) = t/\sqrt{1-t^2}$ for $0 < t < 1$ and that $E(U) = \pi/4$.

58. Let L be the length of the x-coordinate of a randomly oriented unit vector in 2-space. Show that L has probability density $f(x) = 2/(\pi\sqrt{1-x^2})$ for $0 < x < 1$ and that $E(L) = 2/\pi$.

59. (Feller) Why are two violins twice as loud as one? This may sound facetious at first until one recalls that loudness is proportional to the square of the amplitude of the vibration. The incoming waves may be represented by random unit vectors, the length being the amplitude and the angle the phase. When two violins are played, the resulting vector is the vector sum of the two vectors, but since they come from different sources we may regard them as being independent random vectors. Show that the expected value of the square of the length of the sum of the two vectors is twice the expected value of the square of the length of one of them.

60. An isosceles triangle is formed by a unit vector in the x-direction (i.e., in 2-space either (1,0) or (-1,0)) and another in a random direction. Find the distribution of the length of the

third side. Do this both in 2-space and in 3-space.



(2-space)

61. What is the probability that a random quadratic polynomial, $ax^2 + bx + c$, has real roots. Here the coefficients are independent and uniformly distributed on $[0,1]$.

62. A needle of length ℓ is dropped on a grid ruled in a checkerboard pattern with rulings spaced a units apart. What is the average number of lines the needle crosses?

63. A planet contains five small islands which we may regard as five independent random points on a sphere. What is the probability that at least four lie in the same hemisphere?

64. Let P and Q be two independent random points on a circle whose center is 0. What is the distribution of the angle POQ? Do the same for two points on a sphere.

Fluctuation Theory

65. (Epstein) A gambling house offers the following game. After paying an entrance fee E, a coin is tossed until the number of heads exceeds the number of tails. The player is then paid the number of dollars equal to the number of times the coin was tossed. What is the average amount of money that the player expects to receive? [Answer: infinite]

66. A random walk in two or more dimensions is simply two or more independent one-dimensional random walks acting simultaneously. What is the probability that a two-dimensional random walk, starting from the origin, eventually returns to the origin? If it returns, how long, on the average, does it take to do so? Do the same for a three-dimensional random walk. [Answers: probability 1, infinitely long time, probability about 0.239].

67* If the gambling house in exercise 65 above has only N dollars available for winners, what is a fair entrance fee E for the game described there?

Supplementary Exercises

68. Rewrite the following vague conversation using the language of probability theory. You may assume that it is possible to distinguish between "good weather" and "bad weather" unambiguously.

"The Weather Bureau isn't always right, but I would say that they are right more often than not," said Alice thoughtfully.

"Ah, but what comfort is it during miserable weather to know that the forecast was right? If it's wrong that isn't going to affect the likelihood of good weather," retorted Bob.

"You may be right, but that doesn't contradict what I said, even though the forecast is pessimistic only about twice a week," answered Alice persuasively.

Chapter VI The Poisson Process

The Poisson process is the third basic stochastic process, the first two being the Bernoulli and Uniform processes. It can be defined in many ways. We will start with a more abstract approach in section 1. In this section we concentrate on some of the random variables occurring in this process. In the next two sections we give a more intuitive development of the Poisson process based on what we already know about the Uniform and Bernoulli processes. Once we have thoroughly established the properties of the Poisson process, we then turn things around by showing that the Uniform process can be obtained by conditioning the Poisson process!

1. Continuous Waiting Times

Suppose we toss a coin k times and that we get k tails. It is intuitively obvious that on the next toss there is the same probability for heads as ever: the coin does not remember what took place in the past. We can express the fact that a coin has no memory in terms of the single waiting time W_1 as follows

$$P(W_1>k+n|W_1>k) = P(W_1>n).$$

The probability that one will get a run of k+n tails given that one has just gotten a run of k tails is simply the probability that one will get the additional n tails: the preceding tails neither help nor hurt. How long one must wait does not depend on how long one has already waited.

In real life if one is waiting for an incident to take place there is no abstract entity flipping an abstract coin during small discrete time intervals determining when the incident is to occur. For example one might be standing next to a Geiger counter waiting for a click. The waiting time in this case is continuous, but like the Bernoulli waiting time the waiting time has no memory. We express this using conditional probability.

Definition. A positive continuous random variable W is said to have the exponential distribution when

$$P(W>t+s|W>s) = P(W>t)$$

for all positive t,s.

We will also call W a continuous memoryless waiting time, although we will see that the value of W need not represent time. The exponential distribution is an ubiquitous distribution appearing in the most unexpected places.

What is surprising about random variables having the exponential distribution is that the seemingly innocuous assumption we have made in defining this concept determines the probability distribution of W but for a single parameter. To see this let $G(t) = P(W>t)$. The condition $P(W>t+s|W>s) = P(W>t)$ may also be written $P((W>t+s)\cap(W>s)) = P(W>t)P(W>s)$. But the event $(W>t+s)$ is a subevent of the event $(W>s)$. Therefore $(W>t+s)\cap(W>s) = (W>t+s)$. So we may equally well characterize a continuous memoryless waiting time by the condition:

$$P(W>t+s) = P(W>t)P(W>s) ,$$

or in terms of G:

$$G(t+s) = G(t)G(s).$$

From this equation alone we can, using calculus, deduce that $G(t) = Ke^{Ct}$ for suitable constants K and C. Those who have seen this done before can skip the next paragraph.

If we think of G(t+s) as a function of two variables t and s, we may compute the partial derivatives by the chain rule of calculus:

$$\frac{\partial}{\partial t} G(t+s) = G'(t+s)\cdot\frac{\partial(t+s)}{\partial t}$$

$$= G'(t+s)$$

Similarly, $\frac{\partial}{\partial s} G(t+s) = G'(t+s)$. Next we differentiate G(t)G(s) with respect to both t and s:

$$\frac{\partial}{\partial t} (G(t)G(s)) = G'(t)G(s)$$

$$\frac{\partial}{\partial s} (G(t)G(s)) = G(t)G'(s)$$

Therefore

$$G'(t)G(s) = G'(t+s) = G(t)G'(s).$$

Divide both sides by G(t)G(s) to get

$$\frac{G'(t)}{G(t)} = \frac{G'(s)}{G(s)}$$

Now this must hold no matter what t and s are. Therefore

$$\frac{G'(t)}{G(t)} = C$$

for some constant C . Finally if we integrate both sides we get

$$\ell n|G(t)| = Ct + D$$

for some constant D , and this is the same as

$$G(t) = Ke^{Ct}$$

for some constant K.

The distribution of W is therefore

$$F(t) = P(W\leq t) = 1-G(t) = 1-Ke^{Ct} \quad \text{for } t>0.$$

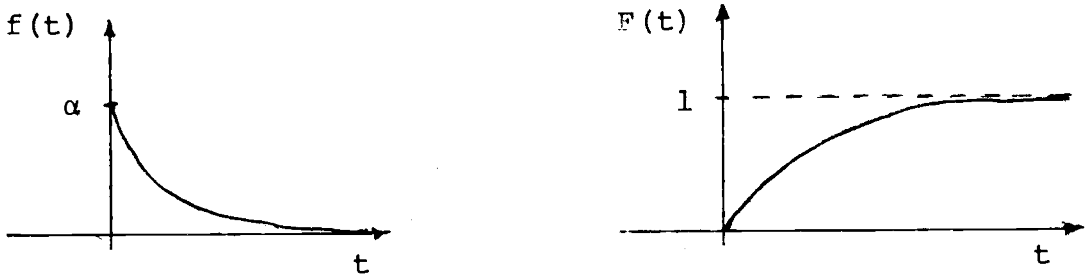
Since we must have $\lim_{t\to\infty} F(t) = 1$, the constant C must be negative. Write $\alpha = -C$. Since we must also have $\lim_{t\to\infty} F(t) = \lim_{t\to 0} F(t) = 0$, the constant K must be 1. We conclude that the probability distribution of a continuous waiting time W is

$$F(t) = P(W\leq t) = 1-e^{-\alpha t} ,$$

for some positive constant α; and its density is

$$f(t) = \alpha e^{-\alpha t}$$

We now see why we say that W is exponentially distributed.



The density and distribution of a continuous waiting time.

The parameter α may be interpreted as the frequency of the incidents in time: roughly speaking, α incidents occur per unit time "on the average."

The power of probabilistic reasoning (made rigorous by conditional probability) is that we may compute the distribution of a random variable without referring to a sample space or to events of it. The distribution is defined purely in phenomenological terms, i.e., in terms of the observed phenomena only.

Consider for example that we have a collection of points dropped independently and uniformly throughout the entire infinite plane. By "uniformly" we mean that the probability of finding a point in a region of finite area t depends only on the area t (not on its shape or location). By "independently" we mean that for two disjoint regions the probability of finding a point in one region is independent of finding a point in the other. Write P(T>t) for the probability of finding no point in a region of area t. Then

$$P(T>t+s) = P(T>t)P(T>s).$$

Therefore, as above,

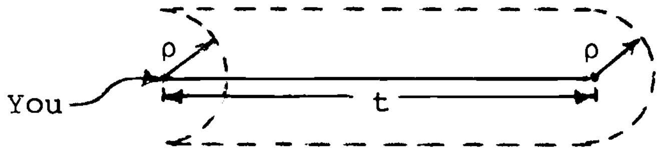
$$P(T>t) = e^{-\alpha t} ,$$

where α may be interpreted as the density of the points dropped on the plane. It is reasonable to regard T as a "waiting area", i.e., "how large an area must a region be in order to find a point in the region?"

Consider next a collection of stars distributed at random in a large region of space. How far away is the nearest neighbor to a star in this region? This is quite similar to the above problem, but we now have three dimensions. Instead of a region of some area t, we use a spherical volume of radius r whose center is the given star. If the average density of the stars is α, then

$$P(\text{Nearest neighbor is more than } r \text{ units away}) = e^{-\frac{4}{3}\pi r^3\alpha}.$$

Suppose we are in a forest with randomly located trees. How far can one see if one looks in one particular direction? By symmetry one may assume that one looks along the positive x-axis from the origin. Assume also, for simplicity, that the trees are all ρ units in radius. Let T be the random variable "how far can one see along the x-axis?" If T is larger than t , then there are no centers of trees in the region indicated below:



The identation on the left side is a consequence of the fact that one happens to be standing at that point. The area of the dotted region is the same as that of a rectangle of sides t and 2ρ. Therefore

$$P(T>t) = e^{-2\alpha\rho t}$$

Needless to say this is an idealized model (trees do not all have the same radius), but it illustrates the basic **idea**

One gets a very similar model when one studies the effect of a beam of high energy protons entering a detector consisting of a tank of liquid hydrogen. Here the "trees" are the nuclei of the hydrogen atoms, although we should build the model in three dimensions instead of two.

To summarize, every "waiting time" for which the future does not depend on the past exhibits exponential decay.

The Gamma Distribution

We just saw that the analog for continuous random variables of the random variable W_1 in the Bernoulli process is an exponential random variable. We now ask for the analog of the k^{th} waiting time W_k in the Bernoulli process. Now the k^{th} waiting time is the sum of the first k gaps in the Bernoulli process: $W_k = T_1 + T_2 + \cdots + T_k$; and the gaps are independent. Therefore we could have computed the distribution of W_k by convolving the distributions of the gaps, all of which are geometric random variables. As an example, the

distribution of W_2 is then the convolution of the distribution of T_1, $q^{n-1}p$, with itself, i.e.

$$P(W_2=n) = \sum_{k=1}^{n-1} q^{(n-k)-1}pq^{k-1}p = \sum_{k=1}^{n-1} q^{n-2}p^2 = (n-1)\cdot q^{n-2}p^2 .$$

Consider now the continuous analogue of a waiting time: the exponential distribution. The sum of two independent exponentially equidistributed random variables T_1 and T_2 may be regarded as the waiting time for the second occurrence, $W_2 = T_1+T_2$, just as in the Bernoulli process. In the next chapter we shall build a more concrete model on which to define this random variable . Although we haven't yet defined the T_i's on a specific sample space, we can nevertheless compute the distribution of the continuous waiting time W_2. It is the convolution of $\alpha e^{-\alpha t}$ with itself:

$$\text{dens}(W_2=t) = \int_0^t \alpha e^{-\alpha(t-s)} \alpha e^{-\alpha s}\, ds$$

$$= \alpha^2 \int_0^t e^{-\alpha t}\, ds$$

$$= \alpha^2 e^{-\alpha t} \int_0^t ds$$

$$= \alpha^2 t e^{-\alpha t}$$

More generally, the density of the kth waiting time W_k is the convolution

$$\underbrace{\alpha e^{-\alpha t} * \alpha e^{-\alpha t} * \ldots * \alpha e^{-\alpha t}}_{\text{k times}} \quad .$$

W_k is the sum $T_1+T_2+\ldots+T_k$ of k independent exponentially distributed random variables, all with parameter α. This convolution is easily computed:

$$\text{dens}(W_k=t) = \frac{\alpha^k t^{k-1}}{(k-1)!} e^{-\alpha t} \quad .$$

We call this the Gamma Distribution. Notice that it has two parameters: α and k.

We end this section by computing the means and variances of the continuous waiting times. It is an easy exercise to verify that if T is exponentially distributed, then $E(T) = 1/\alpha$. where α is the parameter defining the distribution of T. This coincides with our intuitive feeling that α is an "intensity."

The variance of T is easily computed by using integration by parts twice.

$$\begin{aligned}
\mathrm{Var}(T) &= E(T^2) - E(T)^2 \\
&= \int_0^\infty t^2 \, \alpha e^{-\alpha t} \, dt - (\tfrac{1}{\alpha})^2 \\
&= \alpha \Big[-\frac{t^2}{\alpha} e^{-\alpha t} - \frac{t}{\alpha^2} e^{- \ t} + \frac{2}{\alpha^3} e^{- \ t} \Big]_0^\infty - \frac{1}{\alpha^2} \\
&= \frac{2\alpha}{\alpha^3} - \frac{1}{\alpha^2} \\
&= \frac{1}{\alpha^2}
\end{aligned}$$

So the standard deviation $\sigma(T)$ is the same as the mean $E(T)$: both are $1/\alpha$.

Because the k^{th} waiting time W_k of the Poisson process is the sum of k independent, equidistributed exponential random variables, the variance of W_k is $k/_{\alpha}2$.

2. Comparing the Bernoulli and Uniform Processes

We could at this point simply define the Poisson process to be a sequence of independent, exponentially distributed random variables, having the same parameter α. But we prefer to take a different approach, which builds on the two processes we have already thoroughly studied. We therefore now give a detailed comparison of the Bernoulli and Uniform processes emphasizing their similarities and their differences. The Poisson process will be a "limit" of both processes so that, in a sense, it furnishes a formal link between them. In so doing we will discover some new aspects of both these processes.

Parameters

The Bernoulli process depends on a single parameter: the bias p of the coin. The uniform process depends on two parameters: the length a of the interval and the number n of points sampled. There is already a certain asymmetry here. The number of points per unit interval, $\alpha = \frac{n}{a}$, is called the intensity of the uniform process. Different uniform processes having the same intensity are quite similar, especially when n is large.

Sample Spaces

The sample space Ω of the Bernoulli process is the set of all sequences of zeroes and ones. To every such sequence we can associate a set of natural numbers: the set of positions having ones. For example,

$$(0,1,1,0,1,1,0\ldots) \text{ corresponds to } \{2,3,5,6,\ldots\}.$$

This gives us a new way of looking at Ω . It is the set of all subsets of the natural numbers.

The sample space Ω of the uniform process is the set of all sequences $(x_1,x_2,\ldots,x_n)$ of real numbers such that $0 \leq x_i \leq a$. There seems to be little similarity between this sample space and the Bernoulli sample space.

Events

The elementary events of the Bernoulli process are the subsets $H_n = (X_n=1)$ = "the n^{th} toss is heads." The elementary events of the uniform process are the subsets $(s \leq X_i < t)$ = "the i^{th} point falls in $[s,t)$". In both cases the events in general are obtained by intersections, complements and unions from the elementary events.

Random Variables

Up to now we have viewed the random variables X_n of the Bernoulli process and X_i of the uniform process as being fundamental. But there is an alternative point of view. We could equally well define the Bernoulli process by the random variables S_n, the number of successes in the first n tosses. We should write this $S_n^{(p)}$ to denote the fact that it depends on the parameter p. We know that $S_n^{(p)}$ has binomial distribution with bias p:

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

Similarly we could define the uniform process by the random variables $U_{n,a}(t)$, the number of points falling in the interval $[0,t)$. These are random variables we have not yet seen. For each t in $[0,a]$, $U_{n,a}(t)$ is an <u>integer</u> random variable. In fact, one can easily see that $U_{n,a}(t)$ has the binomial distribution, for points fall in $[0,t)$ or in $[t,a]$ with the same probability as a tossed coin with bias $p = \frac{t}{a}$ lands heads or tails, respectively. Therefore,

$$P(U_{n,a}(t) = k) = \binom{n}{k}\left(\frac{t}{a}\right)^k\left(1-\frac{t}{a}\right)^{n-k} .$$

We shall abbreviate $U_{n,a}(t)$ to simply $U(t)$. For each t, $U(t)$ is a new random variable. When a collection of random variables depend on a continuous parameter, we call the collection a random function. Be careful not to think of this as a "randomly chosen function" any more than a random variable is a "randomly chosen variable."

Next we compare the waiting time W_k for the k^{th} success in the Bernoulli process with the k^{th} order statistic $X_{(k)}$. If we think of $[0,a]$ as part of a time axis, there is clearly an analogy between these two random variables. Compare their distribution and density:

$$P(W_k=n) = \binom{n-1}{k-1}q^{n-k}p^k$$

$$\text{dens}(X_{(k)}=t) = \binom{n-1}{k-1}\frac{n}{a}\left(\frac{t}{a}\right)^{k-1}\left(1-\frac{t}{a}\right)^{n-k} .$$

These are quite similar indeed except that in the latter case a factor of $\frac{t}{a}$ has become $\frac{n}{a}$, as a result of differentiation.

We finally come to the gaps in these two processes. In the Bernoulli process, the gaps T_i are equidistributed with geometric distribution:

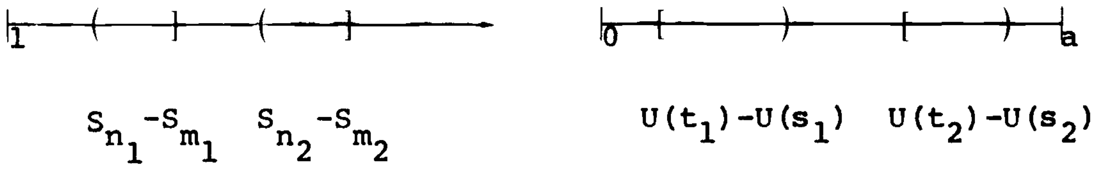
$$P(T_i = k) = q^{k-1}p.$$

The gaps L_i of the uniform process are also equidistributed, having the Dirichlet distribution:

$$\text{dens}(L_i = t) = \frac{n}{a}(1-\frac{t}{a})^{n-1} .$$

The analogy between these two cases is quite striking.

However, the analogy breaks down because the gaps T_i are independent, whereas the gaps L_i are not. To be sure the L_i try as hard as they can to be independent — they are exchangeable — but this is not enough. Another way to see the difference between the two processes is to return to the "fundamental" random variables S_n and $U(t)$. The difference S_n-S_m is the number of successes between m and n. Similarly $U(t)-U(s)$ is the number of points falling between s and t. Now if $(m_1,n_1]$ and $(m_2,n_2]$ are disjoint intervals



of integers, then the random variables $S_{n_1}-S_{m_1}$ and $S_{n_2}-S_{m_2}$

are independent. But even if $[s_1,t_1)$ and $[s_2,t_2)$ are disjoint subintervals of $[0,a]$, $U(t_1)-U(s_1)$ and $U(t_2)-U(s_2)$ are not independent.

The difficulty stems from the fact that the uniform process is on a finite interval and a finite number of points whereas the Bernoulli process is not limited in this way: [Whether we choose to limit the Bernoulli process to a given finite number of tosses is irrelevant, for we can always continue it if we wish. The uniform process has no such option. We can always drop more points, but we cannot extend the interval to a longer one without totally altering our process.]

These considerations suggest that there is a third process that makes the analogy perfect. Just letting the length go to infinity doesn't work because we cannot make sense of sampling a single point or a finite number of points uniformly from an infinite interval. We would like to have a process that is both uniform on an infinite interval and samples an infinite number of points. This works provided we let a and n go to infinitely simultaneously while keeping the intensity $\alpha=\frac{n}{a}$ fixed. Intuitively, because the number of points per unit interval remains the same, in the limit one will have the same intensity, and the uniform processes will converge to a new process.

For example, consider the density of a gap as a and n become large:

$$\text{dens}(L_i=t) = \frac{n}{a}(1-\frac{t}{a})^{n-1}$$

$$= \frac{\alpha(1-\frac{t\alpha}{n})^{n}}{(1-\frac{t\alpha}{n})} .$$

We know from calculus that $\lim\limits_{n\to\infty} (1+\frac{x}{n})^{n} = e^{x}$. Therefore if we let n tend to ∞, the above expression becomes $\frac{\alpha e^{-t\alpha}}{(1-0)} = \alpha e^{-\alpha t}$. That is, in the limit the gaps become exponentially distributed. This is exactly what we would expect, for the gaps of the Bernoulli process are waiting times, and we would hope that the gaps of the new process will be continuous waiting times.

Consider another example: the joint density of two gaps. In the uniform process this density is not the product of the individual densities. In the limit, however, the joint density of two gaps is the product

$$\text{dens}((L_1=t_1)\cap(L_2=t_2)) = \text{dens}((X_{(1)}=t_1)\cap(X_{(2)}=t_1+t_2))$$

$$= \frac{n(n-1)}{a^2}\left(1 - \frac{t_1+t_2}{a}\right)^{n-2}$$

$$= \frac{\alpha(\alpha-\frac{1}{a})\left(1-\frac{\alpha(t_1+t_2)}{n}\right)^n}{\left(1-\frac{\alpha(t_1+t_2)}{n}\right)^2}$$

$$\longrightarrow \frac{\alpha^2 e^{-\alpha(t_1+t_2)}}{(1-0)^2} = \left(\alpha e^{-\alpha t_1}\right)\left(\alpha e^{-\alpha t_2}\right)$$

of the densities. So in the new process the gaps are independent and equidistributed, just as in the Bernoulli process. This helps to confirm our feeling that this is the correct approach.

As a final example, we consider the limit of the random function $U_{n,a}(t)$ as $a,n\to\infty$. Recall that

$$P(U_{n,a}(t) = k) = \binom{n}{k}\left(\frac{t}{a}\right)^k\left(1-\frac{t}{a}\right)^{n-k}$$

$$= \frac{(n)_k}{k!}\left(\frac{\alpha t}{n}\right)^k\left(1-\frac{\alpha t}{n}\right)^{n-k} .$$

As in the last two examples, we can see that $\left(1-\frac{\alpha t}{n}\right)^{n-k} \to e^{-\alpha t}$ because k is a fixed integer. The limit of the first two factors is a bit harder. First write them as: $\frac{(n)_k}{k!} \cdot \frac{(\alpha t)^k}{n^k}$. Then interchange denominators to get: $\frac{(n)_k}{n^k} \cdot \frac{(\alpha t)^k}{k!}$. The second factor is now independent of n and a. The first factor is:

$$\frac{n(n-1)\ldots(n-k+1)}{n \cdot n\ \ldots\ n} = 1\cdot\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{k-1}{n}\right) .$$

Each of these factors approaches 1 as $n\to\infty$. Since there are a fixed number of them, their product approaches 1 as $n\to\infty$. Therefore,

$$\lim_{\substack{n,a\to\infty \\ \frac{n}{a}=\alpha}} P(U_{n,a}(t) = k) = \frac{(\alpha t)^k}{k!}\, e^{-\alpha t}$$

Since $U_{n,a}(t)$ is the fundamental random function of the uniform process, its limit will be the fundamental random function of the new process. We shall write $N_\alpha(t)$ or $N(t)$ for this limit:

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

Notice that the distribution of $N_\alpha(t)$ depends only on the product αt. We call this distribution the Poisson distribution. More precisely:

Definition. An integer random variable X is said to have the Poisson distribution with parameter λ if

$$P(X = k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & \text{if } k \geq 0 \\ 0 & \text{if } k<0. \end{cases}$$

The expectation of such a random variable is

$$E(X) = \sum_k k\, P(X = k)$$

$$= \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda}$$

$$= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}$$

$$= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

$$= \lambda e^{-\lambda} e^{\lambda}$$

$$= \lambda \quad .$$

Therefore, $E(N_\alpha(t)) = \alpha t$.

If we imagine that an infinite number of points are spread on the interval $[0,\infty)$ with density α, then $N(t)$ is the number of points that have fallen in the interval $[0,t)$. The average number of points that fall in $[0,t)$ is $E(N\ (t)) = \alpha t$, and the average number of points per unit interval is $\frac{E(N(t))}{t} = \alpha$. This justifies calling α the density or intensity of the process.

Notice that we may no longer speak of which point has fallen first, which is second and so on. If we return to the uniform process for a moment, we can see why this would be so. Originally we used $X_1,X_2,\ldots,X_n$ as the defining random variables of the uniform process. If we use the random function $U_{n,a}(t)$, we can no longer distinguish which point is which. All we know is which point is $X_{(1)}$, which is $X_{(2)}$, and so on. For example $X_{(1)} = t$ if $U_{n,a}(t) = 0$ and $U_{n,a}(s) = 1$ for $s>t$. We can recover the entire uniform process if we know all the random functions $U_{1,a}(t)$, $U_{2,a}(t), \ldots, U_{n,a}(t)$. But when we let $a,n \to \infty$ we only use the random function $U_{n,a}(t)$. As a result the order statistics make sense in the Poisson process, but there is no analogue of the random variables X_i of the uniform process.

3. The Poisson Sample Space

So far we have discussed the Poisson process from two points of view. We first considered it purely phenomenologically via the random variables T_i and W_k. Next we considered it as the limit of uniform processes as the length of the interval increases. We must reconcile these two approaches, and we do so by an explicit construction of a model.

The Poisson sample space is $\Omega = \{\text{all rare sequences}\}$, where a rare sequence is a set of points in $[0,\infty)$ such that every finite interval has at most finitely many points of the rare sequence, i.e. the sequence doesn't cluster. To avoid confusion the points of a rare sequence are called incidents or blips. Don't confuse this notion with the concept of a sample point. The sample points of Ω in this case are the rare sequences (not the blips).

Defining Ω this way seems very natural because the Poisson process is the limit of the uniform process (of intensity α) on $[0,a]$ as $a\to\infty$. Unfortunately we have allowed n to approach infinity as well. So the intensity is not an intrinsic part of the Poisson sample space as it was for the uniform sample space, where the intensity is $\frac{n}{a}$. The same situation occurs for the Bernoulli process. There the sample space is the same whatever the bias of our coin. It is only through the definition of the probability P that we can say that "the average number of heads in n tosses is np." In a similar way, we shall define a probability P on the Poisson sample space so that the average number of points falling on any interval of length t is αt.

We define the probability P on Ω by means of the random function $N(t)$. We already saw in the last section what the distributions of the random variables $N(t)$ ought to be. We will see in a moment how this point of view leads immediately to probabilities on the elementary Poisson events. All the distributions of the other random variables on Ω will be derived from the distributions of the $N(t)$. We make three fundamental assumptions:

(1) For every nonnegative real number t, N(t) is a nonnegative integer random variable whose value is the number of blips in the interval [0,t). More generally for $s \leq t$, $N(t)-N(s)$ is also an integer random variable whose value is the number of blips in the interval [s,t).

(2) $$P(N(t) - N(s) = k) = \frac{(\alpha(t-s))^k}{k!} e^{-\alpha(t-s)}, \text{ if } 0 \leq s \leq t .$$

That is, on any subinterval [s,t) of $[0,\infty)$, the number of blips occurring has a Poisson distribution. Notice that we assume the density of the blips is independent of the location of the subinterval. In particular,

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

(3) If $[s_1,t_1)$ and $[s_2,t_2)$ are disjoint subintervals of $[0,\infty)$, then $N(t_1)-N(s_1)$ and $N(t_2)-N(s_2)$ are independent random variables. In other words, what happens on disjoint subintervals are independent of one another.

The above three fundamental assumptions implicitly define the events of Ω. Assumption (1) implies that $(N(t)-N(s)=k)$ is an event of Ω for all t and k. We had a different notation for this event in section I.4:

$$\begin{bmatrix} s,\ t \\ k \end{bmatrix} = \{\text{all rare sequences having } k \text{ blips in the interval } [s,t)\}$$
$$= (N(t)-N(s) = k).$$

The above fundamental assumptions may be rewritten in terms of events as follows.

(1) The subsets $\begin{bmatrix} s,\ t \\ k \end{bmatrix}$ of Ω are the <u>elementary Poisson events</u>. An arbitrary event of Ω is obtained from elementary events by intersections, complements and unions.

(2) $$P\left(\begin{bmatrix} s,t \\ k \end{bmatrix}\right) = \frac{(a(t-s))^k}{k!} e^{-\alpha(t-s)}$$

(3) $\begin{bmatrix} s_1,t_1 \\ k \end{bmatrix}$ and $\begin{bmatrix} s_2,t_2 \\ \ell \end{bmatrix}$ are independent events provided that $[s_1,t_1)$ and $[s_2,t_2)$ are disjoint intervals.

Unfortunately there is a problem with the definition of P above. How do we know that conditions (1), (2) and (3) do not imply some subtle contradiction? For the Bernoulli and the uniform processes it was quite obvious that our definition gives a unique value for P(A) no matter how the event A is written in terms of elementary events. Although we have given many reasons for believing that the Poisson process should be well-defined, we haven't proved it yet. [If you are willing to believe that the Poisson process is well-defined, you can skip lightly over the rest of this section.]

For example, consider the event $[{}^{s,t}_{0}]^{c}$; i.e., the event that one or more blips occur in the interval [s,t). We could also write this as:

$$[{}^{s,t}_{0}]^{c} = [{}^{s,t}_{1}] \cup [{}^{s,t}_{2}] \cup \ \ldots$$

The probability of the event on the left hand side above is:

$$\begin{aligned} P([{}^{s,t}_{0}]^{c}) &= 1 - P([{}^{s,t}_{0}]) \\ &= 1 - e^{-\alpha(t-s)} \ . \end{aligned}$$

On the other hand, because the events $[{}^{s,t}_{1}], [{}^{s,t}_{2}], \ldots$ are disjoint, the probability of the right hand side is the following. [Recall that the Taylor series expansion of $e^{\alpha(t-s)}$ is $1 + \alpha(t-s) + \frac{(\alpha(t-s))^2}{2} + \ldots$.]

$$P([{}^{s,t}_{1}]) + P([{}^{s,t}_{2}]) + \ldots$$

$$= \alpha(t-s)e^{-\alpha(t-s)} + \frac{(\alpha(t-s))^2}{2} \ e^{-\alpha(t-s)} + \ldots$$

$$\begin{aligned} &= (e^{\alpha(t-s)} - 1)e^{-\alpha(t-s)} \\ &= 1 - e^{-\alpha(t-s)} . \end{aligned}$$

So we get the same answer either way.

As another example, if $r\leq s\leq t$, then k blips occur in $[r,t)$ if and only if some occur in $[r,s)$ and the rest occur in $[s,t)$. In symbols:

$$\begin{bmatrix} r,t \\ k \end{bmatrix} = \bigsqcup_{\ell=0}^{k} \left(\begin{bmatrix} r,s \\ \ell \end{bmatrix} \cap \begin{bmatrix} s,t \\ k-\ell \end{bmatrix} \right)$$

We can therefore compute $[{r,t \atop k}]$ in two ways. The first way is:

$$P([{r,t \atop k}]) = \frac{(\alpha(t-r))^k}{k!} e^{-\alpha(t-r)} .$$

The second way is the following. It is quite complicated and requires all of our assumptions on P.

$$\begin{aligned} P([{r,t \atop k}]) &= P\left(\bigsqcup_{\ell=0}^{k} [{r,s \atop \ell}] \cap [{s,t \atop k-\ell}]\right) \\ &= \sum_{\ell=0}^{k} P([{r,s \atop \ell}])P([{s,t \atop k-\ell}]) \\ &= \sum_{\ell=0}^{k} \frac{(\alpha(s-r))^\ell}{\ell!} e^{-\alpha(s-r)} \frac{(\alpha(t-s))^{k-\ell}}{(k-\ell)!} e^{-\alpha(t-s)} \\ &= \sum_{\ell=0}^{k} \frac{\alpha^\ell}{\ell!} (s-r)^\ell \frac{\alpha^{k-\ell}}{(k-\ell)!} (t-s)^{k-\ell} e^{-\alpha(t-r)} \end{aligned}$$

$$= \alpha^k \left(\sum_{\ell=0}^{k} \frac{1}{\ell!(k-\ell)!} (s-r)^\ell (t-s)^{k-\ell} \right) e^{-\alpha(t-r)}$$

$$= \frac{\alpha^k}{k!} \left(\sum_{\ell=0}^{k} \frac{k!}{\ell!(k-\ell)!} (s-r)^\ell (t-s)^{k-\ell} \right) e^{-\alpha(t-r)}$$

$$= \frac{\alpha^k}{k!} \left(\sum_{\ell=0}^{k} \binom{k}{\ell} (s-r)^\ell (t-s)^{k-\ell} \right) e^{-\alpha(t-r)}$$

$$= \frac{\alpha^k}{k!} ((s-r) + (t-s))^k e^{-\alpha(t-r)}$$

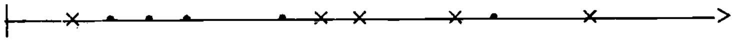
$$= \frac{\alpha^k (t-r)^k}{k!} e^{-\alpha(t-r)} .$$

Notice our use of the Binomial formula. In any case, we again get the same answer either way.

Are there other relations among the events $[{}^{s,t}_{k}]$ not obtainable from the above two examples? The answer is no, but this is not easy to prove: we leave this as an exercise. In any case we have now shown that P is consistently defined. One can in fact show that in some sense every possible probability on the Poisson sample space is a perturbation of the probability P we have just defined (for example, α could vary in time).

Sums of Independent Poisson Random Variables

The Poisson process has an important property we now discuss. Imagine that we have two independent Poisson processes of intensities α and β. To distinguish them we color the blips of the first process red and the blips of the second process blue. Now suppose we are color-blind. What we will see is a Poisson process with intensity $\alpha + \beta$. Let $N_{red}(t)$ and $N_{blue}(t)$ be the random functions of the red and blue processes respectively. We are saying that $N_{red}(t) + N_{blue}(t)$ has Poisson distribution with parameter $(\alpha+\beta)t$.



Two independent Poisson processes × = red
• = blue

Let's prove this. Since we assumed $N_{red}(t)$ and $N_{blue}(t)$ are independant, the distribution of the $N_{red}(t) + N_{blue}(t)$ is the convolution of their individual distributions.

$$P(N_{red}(t)+N_{blue}(t)=n) = \sum_{k=0}^{n} P(N_{red}(t)=k)P(N_{blue}(t)=n-k)$$

$$= \sum_{k=0}^{n} \frac{(\alpha t)^k}{k!} e^{-\alpha t} \frac{(\beta t)^{n-k}}{(n-k)!} e^{-\beta t}$$

$$= \frac{t^n}{n!} e^{-(\alpha+\beta)t} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \alpha^k \beta^{n-k}$$

$$= \frac{t^n}{n!} e^{-(\alpha+\beta)t} (\alpha + \beta)^n$$

$$= \frac{((\alpha + \beta)t)^n}{n!} e^{-(\alpha+\beta)t}.$$

The last expression is the Poisson distribution with parameter $(\alpha+\beta)t$. Thus the sum of independent Poisson random variables is again Poisson. This is an important property of the Poisson process, and we will find some very deep applications of it in later sections.

Physical Systems and the Poisson Process

We have already mentioned several examples of exponentially distributed random variables in section 1. The Poisson process is a sequence of independent exponentially distributed random variables so we shouldn't be surprised at its ubiquity.

Geiger Counters

The first example that comes to mind immediately is the sequence of clicks of a Geiger counter. If we are measuring the radiation of a radioactive sample, the clicks are almost the blips of a Poisson process. Of course, we know that the intensity α will gradually decrease as the sample decays. However, if we choose to measure time so that α becomes constant, the Poisson process is an almost perfect model of the physical system. Even if we measure time in the usual units, the model is very close.

Quality Control

Suppose we have a continuous assembly process (say of a rope or wire) and that this process occasionally produces tiny defects randomly on the rope. If we know the length of the rope and the number of defects, then the uniform process is a good model of this system. On the other hand if we know only the average number of defects per unit length (from prior experience), then the Poisson process is a better model. Even if we also know the length of the rope to be produced, the Poisson process is the better model.

One can use such models for Quality Control. If a long length of rope is being produced, one can sample portions of the rope to determine if the number of defects per unit length is exceeding a specified level of acceptance, as might happen if an assembly machine is out of adjustment.

For example, if the average density of the defects on the rope is $\frac{1}{10}$ defect/foot, then the probability of no defects on a rope of length 10 feet is $e^{-10\cdot\frac{1}{10}} = e^{-1}$. The probability of exactly two defects is $\frac{(\frac{1}{10}\cdot 10)^2}{2} e^{-\frac{1}{10}\cdot 10} = \frac{1}{2} e^{-1}$.

Blips from Space

Suppose one is aiming a radio telescope toward one direction in the sky. The signal one is receiving is a sequence of irregularly spaced radio bursts or "blips". Is the signal simply noise or is it a broadcast from some station? By comparing statistical properties of the blips with the known properties of the Poisson process one can distinguish random noise from a broadcast signal with a certain probability of error.

Seeds on a Cornfield

There is a two-dimensional model analogous to the Poisson process. For a region A in the plane, let $\mu(A)$ be the area of A. We replace the random function N(t) by a random function N(A) = number of blips in the region A. The fundamental assumptions on N(A) are:

(1) For every region A, N(A) is an integer random variable, the number of blips occurring in the region.

(2) $P(N(A) = k) = \frac{(\alpha\mu(A))^k}{k!} e^{-\alpha\mu(A)}$, i.e. N(A) has the Poisson distribution with parameter $\alpha\mu(A)$.

(3) If A and B are disjoint regions, then N(A) and N(B) are independent random variables. [Of course, we could produce a model of this kind in any number of dimensions.]

An example of such a process is the process of sprinkling seeds from an airplane randomly onto a field with some intensity α, the average number of seeds per unit area. As another example, we might have stars spread randomly throughout a large volume of space with some average density α of stars per unit volume.

These are just a small selection of an enormous range of examples of the Poisson process occurring in nature. In fact it is the most common of the four basic stochastic processes. We shall now see how the Poisson process can enrich our understanding of the first two stochastic processes; moreover it will give us a powerful tool for computing probability distributions in these processes.

Gaps and Waiting Times

We must now check that in our model the distributions of the gaps and the waiting times correspond to our earlier computations. Consider first the waiting time W_k for the k^{th} blip.

$(W_k \geq t)$ means that the k^{th} blip has not yet occurred by time t, or equivalently that k-1 or fewer blips have occurred in the interval $[0,t)$. In terms of $N(t)$:

$$(W_k \geq t) = (N(t)=0) \cup (N(t)=1) \cup \ldots \cup (N(t)=k-1).$$

The events of the right hand side being disjoint, we may compute:

$$P(W_k \geq t) = P(N(t)=0) + P(N(t)=1) + \ldots + P(N(t)=k-1)$$

$$= e^{-\alpha t} + \alpha t e^{-\alpha t} + \frac{(\alpha t)^2}{2} e^{-\alpha t} + \ldots + \frac{(\alpha t)^{k-1}}{(k-1)!} e^{-\alpha t} .$$

The density of W_k is the derivative $\frac{d}{dt}(P(W_k \leq t)) = \frac{d}{dt}(1-P(W_k > t))$. Since $P(W_k=t) = 0$, $\text{dens}(W_k=t) = -\frac{d}{dt} P(W_k \geq t) = -\frac{d}{dt}(e^{-\alpha t} + \alpha t\, e^{-\alpha t} + \ldots + \frac{(\alpha t)^{k-1}}{(k-1)!} e^{-\alpha t})$. When we differentiate the latter expression, each term except for the first gives rise to two terms.

$$-\text{dens}(W_k=t) = (-\alpha e^{-\alpha t}) + (\alpha e^{-\alpha t} - \alpha^2 t^{e-\alpha t}) + (\alpha^2 te^{-\alpha t} - \frac{\alpha^3 t^2}{2} e^{-\alpha t})$$

$$+\ldots+ (\frac{\alpha^{k-1}t^{k-2}}{(k-2)!} e^{-\alpha t} - \frac{\alpha^k t^{k-1}}{(k-1)!} e^{-\alpha t}).$$

All the terms then cancel except for the last one so that

$$\text{dens}(W_k=t) = \frac{\alpha^k t^{k-1}}{(k-1)!} e^{-\alpha t} \quad ,$$

which is the gamma distribution. In particular, the first waiting time (the first gap) is exponentially distributed.

To show that the gaps are independent and exponentially distributed with parameter α we use the continuous law of successive conditioning and the continuous law of alternatives in exactly the same way that we used the ordinary law of successive conditioning and the ordinary law of alternatives to compute the distributions of the gaps of the Bernoulli process in section V.2. The verification is left as an exercise.

The Uniform Process from the Poisson Process

We built the model of the Poisson process by thinking of what happens to the uniform process as the length of the interval gets larger. We can turn this around; by conditioning the Poisson process to have exactly n blips in the interval [0,a], we get the uniform process.

To see this we compute the conditional probability

$$P(N(t) = k|N(a) = n) = \frac{P((N(t)=k) \cap (N(a)=n))}{P(N(a)=n)} .$$

Now $(N(a)=n)\cap(N(t)=k)$ says that n blips occur in $[0,a)$ and that k of them occur in $[0,t)$. In other words k blips occur in $[0,t)$ and n-k occur in $[t,a)$. These are disjoint intervals; therefore

$$P(N(t)=k|N(a)=n) = \frac{P((N(t)=k)\cap(N(a)-N(t)=n-k))}{P(N(a)=n)}$$

$$= \frac{P(N(t)=k)P(N(a)-N(t)=n-k)}{P(N(a)=n)}$$

$$= \frac{\frac{(\alpha t)^k}{k!} e^{-\alpha t} \frac{(\alpha(a-t))^{n-k}}{(n-k)!} e^{-\alpha(a-t)}}{\frac{(\alpha a)^n}{n!} e^{-\alpha a}}$$

$$= \frac{n!}{k!(n-k)!} \frac{t^k(a-t)^{n-k}}{a^n}$$

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

We recognize this as the distribution of $U_{n,a}(t)$. Since we can express any computation about order statistics in terms of the random function $U_{n,a}(t)$, we can in principle compute anything about order statistics by conditioning the Poisson process.

For example, we can compute the densities of the order statistics without using a limit argument as we did in section III.3.

$$\text{dens}(X_{(k)}=t) = \text{dens}\ (W_k=t|N(a)=n)$$

$$= \frac{\text{dens}((W_k=t)\cap(N(a)=n))}{P(N(a)=n)}$$

If you feel uneasy about the use of a mixture of a density and a probability in the above computation just rewrite it as:

$$\mathrm{dens}(W_k=t)\cap(N(a)=n)) = \frac{d}{dt} P((W_k \le t)\cap(N(a)=n)).$$

The event $(W_k<t)$ is the same as saying at least k blips occur in $[0,t)$, i.e. $(W_k<t) = (N(t)=k)\cap(N(t)=k+1)\cap \ldots$. On the other hand, if we know that the k^{th} blip has occurred at time t, then $(N(a)=n)$ and $(N(a)-N(t)=n-k)$ are the same event. Therefore

$$\mathrm{dens}(X_{(k)}=t) = \frac{\mathrm{dens}((W_k=t)\cap(N(a)-N(t)=n-k))}{P(N(a)=n)}$$

$$= \frac{\mathrm{dens}(W_k=t)\,P(N(a)-N(t)=n-k)}{P(N(a)=n-k)}$$

$$= \frac{\frac{\alpha^k t^{k-1}}{(k-1)!}\, e^{-\alpha t}\; \frac{\alpha^{n-k}(a-t)^{n-k}}{(n-k)!}\, e^{-\alpha(a-t)}}{\frac{\alpha^n a^n}{n!}\, e^{-\alpha a}}$$

$$= \frac{n!}{(k-1)!(n-k)!}\;\frac{t^{k-1}(a-t)^{n-k}}{a^n}$$

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

Notice that when a Poisson process is conditioned by $(N(a) = n)$, the result is always the uniform process of sampling n points from $[0,a]$ no matter what the intensity was in the original Poisson process. This is further confirmation that the Poisson process is the process of sprinkling points at random on $[0,\infty)$.

4. The Schrödinger Method

One of the most striking applications of the Poisson process is to the discrete problem of counting the number of ways of putting balls into boxes. We consider the problem in full generality. That is, we want a technique whereby if we are given k balls and n boxes and if we are given any restrictions whatsoever on the occupation numbers, then we can compute how many ways this can be done. For example, one might restrict each box to contain zero, one or two balls. For small k and n, one could exhaustively enumerate the possibilities. But for k and n even as small as 10, this is already a very non-trivial problem. As another example, suppose we require that if the third box has an odd number of balls then the fifth box has a multiple of seven balls in it. We need a very systematic procedure if we are to give a reasonable solution to such a counting problem. The technique we will develop is due to Schrödinger, and we call it the <u>randomization</u> <u>technique</u> for reasons we will see in the next section.

We begin with a formula from calculus whose significance is seldom made very clear: Taylor's formula. The difficulty stems from a common misconception that one is supposed to use this formula to compute the Taylor expansion of a function. Although in principle this is possible, this is quite misleading. In fact one usually computes the Taylor expansion by some other technique (and there are many such techniques). One uses the Taylor formula to compute the values of the derivatives of the function at 0 rather than the other way around!

Taylor's formula. Every function f that can be differentiated infinitely many times at 0 has a unique power series expansion, called the Taylor expansion of f at 0:

$$f(0) + f'(0)x + \frac{1}{2}f''(0)x^2 + \cdots + \frac{1}{n!}f^{(n)}(0)x^n + \cdots$$

The notation $f^{(n)}(0)$ is an abbreviation for

$$\left[\frac{d^n}{dx^n}f(x)\right]_{x=0}.$$

For example, if one wishes to compute the Taylor expansion of $f(x) = \sqrt{(1+x)}$, one should not start differentiating repeatedly. The best way is to use the binomial expansion:

$$(1+x)^\alpha = 1 + \binom{\alpha}{1}x + \binom{\alpha}{2}x^2 + \cdots\cdot ,$$

where the binomial coefficient $\binom{\alpha}{k} = \frac{(\alpha)_k}{k!}$ makes sense for any number α as we already noted in section III.6. So for example

$$f(x) = \sqrt{(1+x)} = 1 + \binom{1/2}{1}x + \binom{1/2}{2}x^2 + \cdots ,$$

and by Taylor's formula we see that

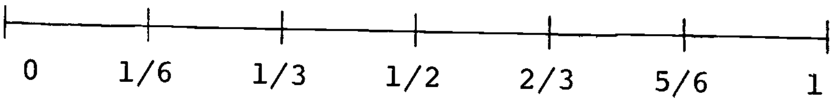
$$\frac{1}{6}f'''(0) = \binom{1/2}{3} = \frac{(1/2)(-1/2)(-3/2)}{3!} = \frac{1}{16} .$$

Suppose that we are given some conditions on the occupation numbers of n boxes. The number of ways we can place k balls into the n boxes subject to these conditions is $n^k P(B_k)$, where B_k is the event:

B_k = "a placement of k balls into n boxes satisfies the given conditions on the occupation numbers."

This event is a subset of the sample space Ω of all placements of k balls into n boxes, each being equally likely.

To compute $P(B_k)$ we construct "physical" boxes from an interval of length $[0,1]$ by cutting it into subintervals each of length $1/n$. Then $P(B_k)$ is the probability that n points dropped



6 boxes made from $[0,1]$

at random uniformly on $[0,1]$ will satisfy the given conditions on the occupation numbers. This converts any balls-into-boxes problem into a computation of the probability of a certain event of the uniform process.

Now comes the important step. We ask a different question. What is the probability that for a rare sequence of blips on $[0,\infty)$ those blips falling in $[0,1]$ satisfy the given conditions? Call this event A. At first it appears that the computation of $P(A)$ is a much harder problem, but we shall see that it is actually much easier. Once we know $P(A)$, we still do not yet know $P(B_k)$ because the events are in completely different processes. However, we just saw that whenever we condition the Poisson process by the event $(N(a) = k)$, the result is a Uniform process. Therefore,

$$P(B_k) = P(A|N(1) = k),$$

and this holds no matter what the intensity α of the original Poisson process. The relationship between $P(A)$ and the condition probabilities $P(A|N(1) = k)$ is given by the law of alternatives:

$$\begin{aligned} P(A) &= \sum_{k=0}^{\infty} P(A|N(1) = k)P(N(1) = k) \\ &= \sum_{k=0}^{\infty} P(B_k)\, \frac{\alpha^k}{k!}\, e^{-\alpha}. \end{aligned}$$

If we multiply both sides of this equation by e^{α}, we get the important formula:

$$P(A)e^{\alpha} = \sum_{k=0}^{\infty} P(B_k)\frac{\alpha^k}{k!} .$$

The left hand side is a function $f(\alpha)$ of the variable α because A is an event of the Poisson process, which depends on the intensity α. The probabilities $P(B_k)$, on the other hand, do not depend on α. Therefore, by the Taylor formula,

$$P(B_k) = \left[\frac{d^k}{d\alpha^k} f(\alpha)\right]_{\alpha=0}$$

or

$$P(B_k) = \left[\frac{d^k}{d\alpha^k}(P(A)e^{\alpha})\right]_{\alpha=0} .$$

Consider the following example. Suppose that the conditions on the occupation-numbers are that $\mathcal{O}_i$ be zero or one for all boxes i. We computed $P(B_k)$ for this case in section III.2: $P(B_k) = \frac{(n)_k}{n^k}$. To compute this using the randomization technique we must first compute $P(A)$, where A is the event "either no blips or just one blip occur in each of the n subintervals of length $1/n$ of $[0,1]$." If we write A_i for the event "box i (i.e. the subinterval $[(i-1)/n, i/n)$ of $[0,1]$ has either no blips or just one blip," then

$$A = A_1 \cap A_2 \cap \cdots \cap A_n .$$

In terms of elementary Poisson events,

$$A_i = \begin{bmatrix} \frac{i-1}{n} , \frac{i}{n} \\ 0 \end{bmatrix} \bigcup \begin{bmatrix} \frac{i-1}{n} , \frac{i}{n} \\ 1 \end{bmatrix} = \left(0 \le N\left(\tfrac{i}{n}\right) - N\left(\tfrac{i-1}{n}\right) \le 1\right).$$

So $P(A_i) = e^{-\alpha/n} + \frac{\alpha}{n} e^{-\alpha/n} = \left(1 + \frac{\alpha}{n}\right)e^{-\alpha/n}$.

Now comes the crucial step. In the Poisson process, the number of blips in disjoint intervals are independent of one another. This is not true for the uniform process, and it is not true for occupation numbers of balls into boxes. It is this fact about the Poisson process that makes this technique so effective. $P(A)$ is easy to compute because the computation of $P(A_i)$ is a routine application of the definition of the Poisson process, and $P(A)$ is the product $P(A_1)P(A_2)\cdots P(A_n)$.

$$P(A) = \underbrace{\left(1 + \frac{\alpha}{n}\right)e^{-\alpha/n}\cdot\left(1 + \frac{\alpha}{n}\right)e^{-\alpha/n}\cdot \;\cdots\; \cdot\left(1 + \frac{\alpha}{n}\right)e^{-\alpha/n}}_{n \text{ factors}}$$

$$= \left(1 + \frac{\alpha}{n}\right)^n e^{-\alpha}.$$

Therefore $P(A)e^{\alpha} = \left(1+\frac{\alpha}{n}\right)^n$ and $P(B_k) = \left[\frac{d^k}{d\alpha^k}\left(1 + \frac{\alpha}{n}\right)^n\right]_{\alpha=0}$.

The formula $P(B_k) = \left[\frac{d^k}{d\alpha^k}\left(1 + \frac{\alpha}{n}\right)^n\right]_{\alpha=0}$ is a perfectly legitimate answer to this problem. We can, however, put this into a nicer form by first expanding $\left(1 + \frac{\alpha}{n}\right)^n$ using the binomial formula and by doing little rearranging:

$$\left(1 + \frac{\alpha}{n}\right)^n = \sum_{k=0}^{n} \binom{n}{k}\left(\frac{\alpha}{n}\right)^k$$

$$= \sum_{k=0}^{n} \frac{(n)_k}{k!}\,\frac{\alpha^k}{n^k}$$

$$= \sum_{k=0}^{n} \frac{(n)_k}{n^k}\,\frac{\alpha^k}{k!}\;.$$

Therefore $P(B_k) = \frac{(n)_k}{n^k}$, exactly as we got previously.

It appears that this is the harder way to compute this answer, but that is only because this example is special. In most cases this technique is considerably easier.

As a harder example consider the event B_k = "every box has at least one ball in it." We computed $P(B_k)$ in section IV.8, using the inclusion-exclusion principle. This was quite an elaborate argument. To compute $P(B_k)$ using randomization, we first compute $P(A)$, where A = "all n boxes of $[0,1]$ have at least one blip." As above, let A_i = "box i has at least one blip." Then

$A_i = \begin{bmatrix} \frac{i-1}{n} , \frac{i}{n} \\ 0 \end{bmatrix}^c$ so that $P(A_i) = 1 - e^{-\alpha/n}$. Hence $P(A) = (1-e^{-\alpha/n})^n$ and $P(A)e^{\alpha} = (1-e^{-\alpha/n})^n e^{\alpha} = (e^{\alpha/n}-1)^n$. Therefore

$$P(B_k) = \left[\frac{d^k}{d\alpha^k}\left(e^{\alpha/n} - 1\right)^n \right]_{\alpha=0} .$$

Now we could leave the answer in this form, but we can derive a better expression by using the binomial formula and the well-known Taylor expansion of the exponential function.

$$\begin{aligned} (e^{\alpha/n} - 1)^n &= \sum_{j=0}^{n} \binom{n}{j} (-1)^j (e^{\alpha/n})^{n-j} \\ &= \sum_{j=0}^{n} \binom{n}{j} (-1)^j e^{(n-j)\alpha/n} \\ &= \sum_{j=0}^{n} \binom{n}{j} (-1)^j \left\{ \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{n-j}{n}\right)^k \alpha^k \right\} \end{aligned}$$

$$= \sum_{j=0}^{n} \binom{n}{j} (-1)^j \left(\sum_{k=0}^{\infty} \left(1 - \frac{j}{n}\right)^k \frac{\alpha^k}{k!} \right)$$

$$= \sum_{j=0}^{n} \sum_{k=0}^{\infty} \binom{n}{j} (-1)^j \left(1 - \frac{j}{n}\right)^k \frac{\alpha^k}{k!}$$

$$= \sum_{k=0}^{\infty} \underbrace{\left(\sum_{j=0}^{n} \binom{n}{j} (-1)^j \left(1 - \frac{j}{n}\right)^k \right)}_{P(B_k)} \frac{\alpha^k}{k!} .$$

Therefore, $P(B_k) = \sum_{j=0}^{n} (-1)^j \binom{n}{j} \left(1 - \frac{j}{n}\right)^k = 1 - \binom{n}{1}\left(1 - \frac{1}{n}\right)^k$ $+ \binom{n}{2}\left(1 - \frac{2}{n}\right)^k \cdots$.

For another example, suppose we want every box to have at most two balls. By now one should be able to compute this immediately:

$$P(A) = \left(e^{-\alpha/n} + \frac{\alpha}{n} e^{-\alpha/n} + \frac{\alpha^2}{2n^2} e^{-\alpha/n}\right)^n$$

$$P(A) e^{\alpha} = \left(1 + \frac{\alpha}{n} + \frac{\alpha^2}{2n^2}\right)^n$$

$$P(B_k) = \left[\frac{d^k}{d\alpha^k} \left(1 + \frac{\alpha}{n} + \frac{\alpha^2}{2n^2}\right)^n \right]_{\alpha=0} .$$

Summary

To compute $P(B_k)$, where B_k is the event "a placement of k balls into n boxes satisfies certain given conditions on the occupation numbers," we do the following:

1. Let A be the event of the Poisson process that "the rare sequence of blips on $[0,\infty)$ has the property that the blips falling in $[0,1]$ satisfy the given conditions on the occupation numbers." Box i is now the subinterval $\left[\frac{i-1}{n}, \frac{i}{n}\right)$.

2. Compute $P(A)$. This will be a function of α. Generally this will not be difficult to compute because what happens in disjoint intervals of the Poisson process are independent of one another.

3. Apply the formula: $P(B_k) = \left[\frac{d^k}{d\alpha^k} P(A)e^{\alpha}\right]_{\alpha=0}$. One can often apply some formulas such as the binomial formula, to expand $P(A)e^{\alpha}$ thereby deriving another expression for $P(B_k)$.

5. Randomized and Compound Processes

Randomization is a general term for the method whereby new stochastic processes are created by allowing a parameter of a given stochastic process to be chosen randomly according to some distribution.

Randomized Uniform Process

For example, consider the uniform process of sampling n points from the interval $[0,a]$. Suppose that instead of sampling a fixed number n of points we sample a random number N of points. That is, we consider the two-step process:

1) Choose a number N of points to be sampled, according to some probability distribution:

$$P(N = n) = p_n$$

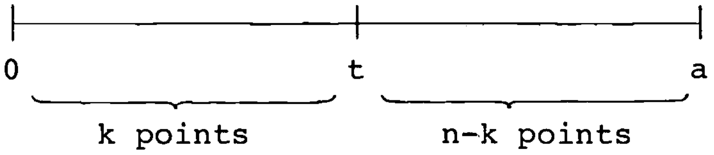
2) Once N is known, sample that many points from $[0,a]$ according to the uniform process.

We have replaced the fixed number n of points by the integer random variable N, having probability distribution $P(N = n) = p_n$. The result is a new stochastic process, the randomized uniform process.

Every question we have asked about the ordinary uniform process can now be asked for the randomized process. To compute the answers in this new process, we use the law of alternatives. For example, we might ask what the probability is for exactly k points to be in the interval $[0,t]$. Call this event $A_{k,t}$. We want to compute $P(A_{k,t})$. By the law of alternatives,

$$P(A_{k,t}) = \sum_{n=0}^{\infty} P(A_{k,t}|N = n)P(N = n).$$

Now $P(A_{k,t}|N = n)$ is the probability that exactly k points are in $[0,t]$ but in the ordinary uniform process of sampling a fixed number n of points from $[0,a]$. We computed this probability back in section VI.1, where we denoted it by $P(U_{n,a}(t) = k)$.



$$P(A_{k,t}|N = n) = P(U_{n,a}(t) = k) = \binom{n}{k}\left(\frac{t}{a}\right)^k\left(1 - \frac{t}{a}\right)^{n-k} .$$

Therefore the unconditional probability $P(A_{k,t})$ in the randomized uniform process is

$$P(A_{k,t}) = \sum_{n=0}^{\infty} \binom{n}{k}\left(\frac{t}{a}\right)^k\left(1 - \frac{t}{a}\right)^{n-k} p_n.$$

If the probabilities p_n have a nice form, then it may be possible to simplify this expression, but normally the answer to a question about a randomized process will be in the form of an infinite series.

There are two reasons why one would randomize a stochastic process. The first is that it allows one to produce more general models of phenomena, which can be more realistic reflections of the phenomena being studied. We shall see examples of these in the exercises. Perhaps more important is the second reason: randomization can be used as a very powerful and effective computational tool. In fact this is one of the most important uses of probability theory. Problems that cannot be solved by direct means can be solved by allowing certain parameters to be random variables. The technique of the last section is just one of these.

Consider another example in the randomized uniform process. Let B_t be the event "in the process of sampling N points uniformly from $[0,1]$, all the points appear in $[0,t]$." Then

$$P(B_t) = \sum_{n=0}^{\infty} P(B_t|N = n)P(N = n),$$

by the law of alternatives. Now $P(B_t|N = n)$ is the probability that all the points of the ordinary uniform process of sampling n points from $[0,1]$ occur in $[0,t]$. Therefore, $P(B_t|N = n) = t^n$, since we have chosen the length of the interval to be 1. Therefore,

$$P(B_t) = \sum_{n=0}^{\infty} t^n p_n.$$

This is a function of t that is usually called the generating function of the sequence $\{p_n\}$.

The technique of generating functions is important both in probability theory and in other branches of mathematics. Unfortunately it is usually defined by fiat with little motivation beyind saying that it is useful. Using probability theory we see more intuitively how it arises. Namely, given a sequence $\{p_n\}$ forming a probability distribution, the generating function $f(t) = \sum_{n=0}^{\infty} p_n t^n$ of the sequence is the probability that a random number of points, the number chosen according to the distribution p_n, when sampled from the unit interval, all occur in the subinterval $[0,t]$. In other words, the generating function is a way of studying a sequence $\{p_n\}$ by setting up a certain experiment using the sequence $\{p_n\}$ and by studying the properties of this experiment. This is the underlying reason why this technique turns out to be so useful.

Randomized Poisson Process

Now consider the Poisson process. Since this process may be regarded as being the uniform process as n and a approach infinity but with $\alpha = \frac{n}{a}$ fixed, we see that the intensity α is the analog in the Poisson process of the number of points sampled in the uniform process. The randomized Poisson process is a Poisson process but with a random intensity A (capital alpha) instead of a fixed intensity. More precisely, this process is again a two-step process:

1) Choose an intensity α according to the density function $g(\alpha) = \mathrm{dens}(A = \alpha)$ of the positive continuous random variable A.

2) Observe a rare sequence of blips in the Poisson process having the chosen intensity.

Let T be the waiting time for the first blip in the randomized Poisson process. To compute the distribution of T, we use the law of alternatives but this time the continuous version.

$$P(T > t) = \int_0^\infty P(T > t | A = \alpha)\,\mathrm{dens}(A = \alpha)\,d\alpha.$$

The conditional probability $P(T > t | A = \alpha)$ is computed in the ordinary Poisson process with intensity α. Therefore:

$$P(T > t) = \int_0^\infty e^{-\alpha t} g(\alpha)\,d\alpha.$$

The function $\hat{g}(t) = \int_0^\infty e^{-\alpha t} g(\alpha)\,d\alpha$ is called the Laplace transform of the function $g(\alpha)$.

The Laplace transform is an important technique in engineering and in the sciences as well as in mathematics. We now see why. If we are given a function $g(\alpha)$ forming the probability density of a positive random variable, we can study $g(\alpha)$ by setting up an experiment and then studying the properties of the experiment. The experiment consists of waiting for the first blip of a Poisson process whose intensity is chosen according to the density $g(\alpha)$. The probability distribution of this experiment is $1 - \hat{g}(t)$, where $\hat{g}(t)$ is the Laplace transform of $g(\alpha)$.

As a simple example of this point of view, we can explain an important property of the Laplace transform: the Laplace transform of the convolution of functions is the product of their Laplace transforms:

$$\widehat{f*g} = \hat{f}\hat{g}.$$

Suppose that $f(\alpha)$ and $g(\alpha)$ are the densities of independent random variables A and B. Their convolution is the density of the sum $A + B$. Let T be the waiting time for the first gap in the Poisson process with random intensity $A + B$. We can compute $P(T > t)$ in two ways. Since the density of $A + B$ is $f*g$, we know that $P(T > t) = \widehat{f*g}$. On the other hand, we may view the event $(T > t)$ in another way. Sprinkle blips on $[0,\infty)$ with intensity A and then with intensity B. Then $(T > t) = (T_A > t) \cap (T_B > t)$, where T_A is the waiting time for the first A-blip and T_B is the waiting time for the first B-blip. Since these two kinds of blips were sprinkled independently, $P(T > t) = P(T_A > t)P(T_B > t) = \hat{f}(t)\hat{g}(t)$. Therefore $\widehat{f*g} = \hat{f}\hat{g}$.

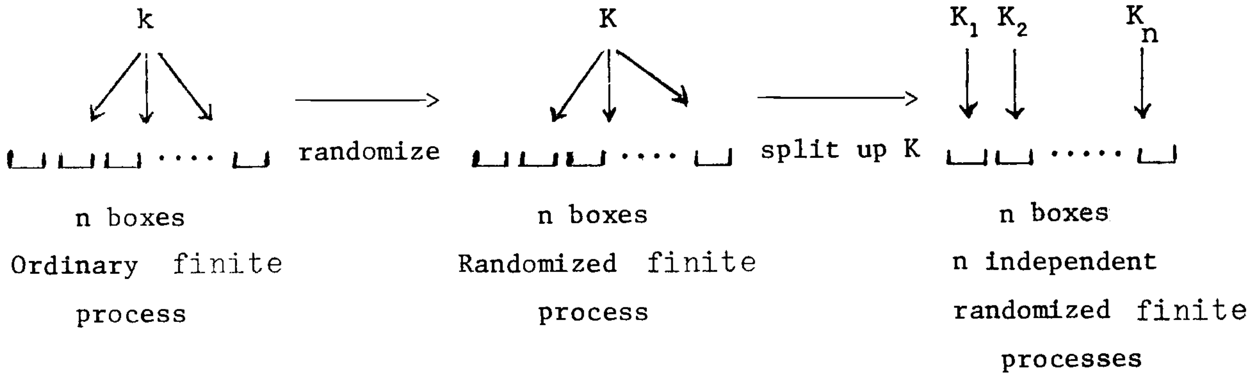
Yoga Randomizing by an integer or a continuous random variable results in a "generating function" or a "transform" of the distribution or density, respectively.

All transforms can be given a probabilistic (possibly quantum probabilistic) interpretation. The Fourier transform is perhaps the deepest example of a transform, since it is intimately connected with quantum mechanics.

Finite Sampling Processes

The finite sampling process or balls-into-boxes can also be randomized. Namely we choose a random number K of balls and then place them randomly into the n boxes. If we can make a judicious choice of distribution for K, then we can possibly make computations in the finite sampling process easier. It was Schrödinger's observation that a good choice for the distribution of K is the Poisson distribution.

The reason that the Poisson distribution works so well is the fact about the Poisson process noted earlier: if we combine two independent Poisson processes with intensities α and β, the result is a Poisson process with intensity $\alpha + \beta$. In terms of the Poisson distribution this says that if X and Y are independent Poisson random variables of parameters λ and μ, then $X + Y$ has Poisson distribution with parameter $\lambda + \mu$. We simply reverse this. Suppose that K has Poisson distribution with parameter α. Then $K = K_1 + K_2 + \cdots + K_n$, where the K_i are independent Poisson random variables each with parameter α/n. The randomized finite sampling process then "splits up" into n independent randomized finite sampling processes. Each of



these randomized finite processes consists of placing a random number K_i of balls into <u>one</u> box. In the last section we used a more specific model. There K was $N(1)$, the number of points occuring in $[0,1)$ in the Poisson process of intensity α. Then $N(1) = N_1\left(\frac{1}{n}\right) + N_2\left(\frac{1}{n}\right) + \cdots + N_n\left(\frac{1}{n}\right)$ is a sum of n independent Poisson random variables each of intensity α/n, where $N_i\left(\frac{1}{n}\right)$ is the number of points occurring in $[(i-1)/n, i/n)$, i.e. $N_i\left(\frac{1}{n}\right) = N\left(\frac{i}{n}\right) - N\left(\frac{i-1}{n}\right)$.

By randomizing, we made non-independent random variables (the occupation numbers) independent. We return to the non-randomized process by conditioning the randomized one, using the law of alternatives.

A generalization of this process immediately comes to mind. We could just as easily drop balls into boxes of different sizes. That is, such that the balls are not equally likely to fall into the various boxes. The Schrödinger technique works just as well in this case; the only change required is that K be split into a sum $K_1+\cdots+K_n$, where K_i is Poisson with parameter $p_i\alpha$, p_i being the probability that any given ball falls in box i. This is the physicists' model of a classical statistical mechanical system. Here p_i is related to the energy of the state represented by box i.

6.* Reliability Theory

Distribution	Type	Parameter(s)	Model(s)
Exponential	continuous	α	W_1 or any T_k in the Poisson process
Gamma	continuous	α, k	W_k in the Poisson process
Poisson	integer	λ	$N(t)$ in the Poisson process, where $\lambda = \alpha t$.

Distribution	Distribution or Density	Mean	Variance
Exponential	$f(t) = \alpha e^{-\alpha t}$	$1/\alpha$	$1/\alpha^2$
Gamma	$f(t) = \frac{\alpha^k t^{k-1}}{(k-1)!} e^{-\alpha t}$	k/α	k/α^2
Poisson	$p_k = \frac{\lambda^k}{k!} e^{-\lambda}$	λ	λ

Table of Poisson Distributions

Fact If N and M are independent Poisson random variables whose parameters are λ and μ, respectively, then $N + M$ is also Poisson but with parameter $\lambda + \mu$.

Bernoulli	Poisson	Uniform
p = bias	α = intensity or average number of blips per unit interval.	n = no. of points sampled. a = length of interval. $\frac{n}{a}$ = intensity.
X_i = outcome of i^{th} toss. independent equidistributed Binomial distribution		$X_i = i^{th}$ point sampled. independent equidistributed Uniform distribution
S_n = no. of successes in first n tosses. Binomial distribution	$N(t)$ = no. of blips in $[0,t)$. Poisson distribution	$U(t)$ = no. of points in $[0,t)$. Binomial distribution
$W_k = k^{th}$ waiting time. Negative binomial distribution	$W_k = k^{th}$ waiting time. Gamma distribution	$X_{(k)} = k^{th}$ order statistic. Dirichlet distribution
$T_i = i^{th}$ gap. independent equidistributed Geometric distribution	$T_i = i^{th}$ gap. independent equidistributed Exponential distribution	$L_i = i^{th}$ gap. not independent exchangeable Dirichlet distribution

Table of Analogies: Bernoulli, Poisson and Uniform Process

7. Exercises for

Chapter VI. The Poisson Process

1. You are the captain of the Bicentennial Eagle, a spaceship that has just returned from hyperspace to ordinary space, only to encounter the debris of a recently destroyed planet. The debris consists of essentially spherical rocks 20m in radius. The destroyed planet was originally the same size as the earth, and its debris is now uniformly scattered throughout a region 10^6km in radius. Your manuvering jets are temporarily out of order. If you are headed directly toward the center of the debris, what are your chances of getting all the way through the debris without a collision? Assume that your ship has a circular cross-section of radius 10m. Explain any assumptions you may be making.

2. At 5×10^4km/hour how long would you have in exercise 1 to repair your manuvering jets before your chances of a collision reach 10%? Explain precisely what you are computing in this problem.

3. A beam of protons is accelerated to high energy and is deflected so that it encounters a pool of liquid hydrogen. The tracks of the protons in the beam are visible in this detector, and one can easily see where a proton in the beam collides with a proton in the pool of liquid hydrogen. Describe how far a given proton travels before it collides with a proton in the pool.

4. There exist enzymes that attack only a certain nucleotide sequence in a chromosome. Describe a means of testing whether or not a given nucleotide sequence appears randomly in a given chromosome.

5. In the birthday coincidence problem (exercise II. 9), the paradox comes from thinking that one is looking for another person with the same birthday as your birthday. Compute the distribution of the number of persons chosen at random you must ask until you find one with the same birthday as yours. What kind of distribution is it? Find an exponential distribution that approximates it. Compute the average value of this random variable as well as the number of persons one must ask in order to have a 50% chance of finding one with your birthday.

6. How does the answer to exercise 5 change if we include February 29th as a possible birthdate?

7. Roughly speaking, the relationship between the birthday problem in exercise 5 and the birthday coincidence problem in exercise II. 9 is that in either case we have a certain number of pairs of persons from which we look for a birthday coincidence, but that in the former problem we consider a collection of pairs all of which have one given person in common whereas in the latter we consider all pairs from a set of persons. For example we saw that in a class of 23 students there is about a 50% chance of a birthday coincidence. Such a class has $\binom{23}{2} = \frac{23 \cdot 22}{2} = 253$ pairs of students. Compare this with the last part of your answer to exercise 5.

8. In a large class the students call out their birthdays until someone in the class finds that his or her birthday has been called. Technically this is not a random variable since it is possible that no pair of students have the same birthday. However, if we assume that a match will eventually be found, then it is a random variable which is approximately exponentially distributed. Find the parameter for this exponential distribution, and compute its mean. Compare with exercise III.25.

9. A sociobiologist wishes to test whether or not birds of a certain species practice territorial spacing of their nest locations. Compute the distribution of the distance of a given nest from its nearest neighbor. Use this to formulate a statistical test.

10.* Let $Y_1, Y_2, \cdots, Y_n$ be n independent, exponentially distributed random variables, each with parameter α. Compute the order statistics $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ of these random variables. One can do this in two ways. Either change variables and convert to a Uniform process (see section V.8) or use a modification of the reasoning used in exercise III.53 which was made rigorous in section V.7 ("needles on a stick problem").

11.* Compute the expectations $E(Y_{(i)})$ in exercise 10 above. Compare with the expectations of the order statistics of the gaps in the Uniform process (exercise III.53).

12.* (Feller) Three persons A, B and C arrive at a post office simultaneously. There are two counters, and these are taken immediately by A and B. Assume the service time of a given individual is exponentially distributed with parameter α. Assume also that different

Chapter VII Entropy and Information

That probability is closely connected with information should come as no surprise after problems such as exercise V.5 (the jailer paradox). What entropy does is to make this connection precise. In section 1 we discuss entropy for finite-valued random variables. In the next section we give a dramatic application of the law of large numbers to information theory: the Shannon Coding Theorem. Finally in section 3 we turn to the case of continuous random variables and prove that essentially all the interesting distributions we have seen in probability theory may be defined by entropy considerations.

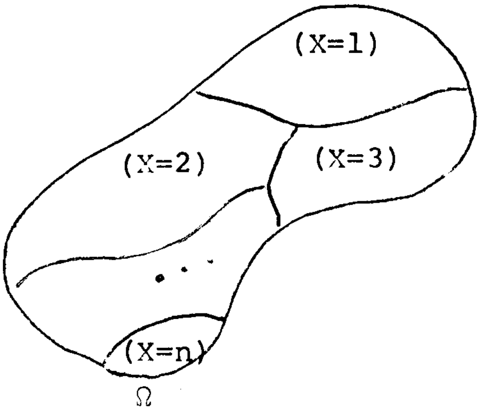
1. Discrete Entropy

We will start by defining entropy for integer random variables taking only finitely many values. Later in a step-by-step procedure, we will extend the concept to continuous random variables.

Partitions.

A random variable is said to be a *finite-valued random variable* if it takes finitely many values. For example, S_n in the Bernoulli process is finite-valued since it can only take on values from 0 to n.

If X is a finite-valued random variable whose values are 1,2,3,...,n, then X determines the events (X=1), (X=2),...,(X=n). Moreover, every outcome of Ω is in exactly one of these events. We call this situation a *partition* of Ω:



In general, a partition π of Ω is a collection of nonempty events $B_1, B_2, \ldots, B_n$ called the blocks of π such that:

(a) no two blocks intersect,

(b) every sample point is in some block, i.e.

$$\bigcup_i B_i = \Omega \ .$$

The only difference between a random variable X and a partition π is that a random variable consists not only of a partition but also of a label (the value X takes on that block) for each block. The partition $\pi(X)$ defined by X is the partition whose blocks are (X=1), (X=2),...,(X=n), i.e. $\pi(X)$ is obtained by ignoring the particular labels that X attaches to the events it defines.

More generally suppose that we have a number of finite-valued random variables $X_1, \ldots, X_r$. The smallest events that one can define by these random variables are the events

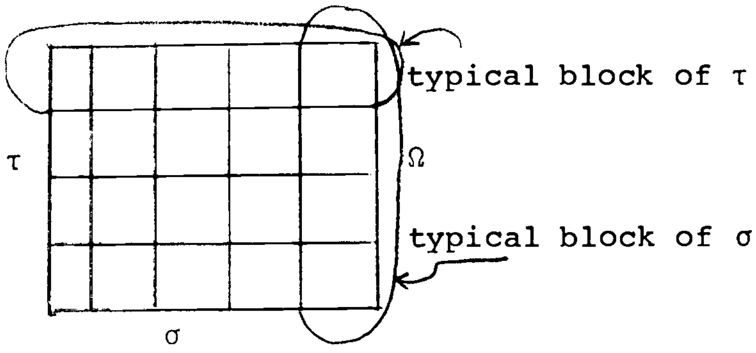
$$(X_1=i_1)\cap(X_2=i_2)\cap\ \ldots\ \cap(X_r=i_r)\ ,$$

and any event definable by the random variables $X_1,\ldots,X_r$ is necessarily a union of some of the above events. The partition whose blocks are the above events is called the (joint) partition $\pi(X_1,\ldots,X_r)$ defined by $X_1,\ldots,X_r$. The partition $\pi(X_1,\ldots,X_r)$ is related to the partitions $\pi(X_1)$, $\pi(X_2),\ldots,\pi(X_r)$ by means of the operation on partitions called the meet. In general if σ and τ are two partitions, whose blocks are $C_1,C_2,\ldots,C_\ell$ and $D_1,D_2,\ldots,D_n$ respectively, then the meet of σ and τ, written $\sigma\wedge\tau$ is the partition whose blocks are $C_i\cap D_j$ whenever they are nonempty. In terms of the meet, $\pi(X_1,X_2,\ldots,X_r) = \pi(X_1)\wedge\pi(X_2)\wedge\ldots\wedge\pi(X_r)$.

As the joint distribution of random variables determines everything about their "correlation" so the joint partition of a set of partitions determines their correlation. In particular, it is easy to see that independence of random variables is really a property of the partitions defined by them. Let σ and τ be two partitions. We say σ and τ are independent if and only if

$$P(C\cap D) = P(C)P(D)$$

for all blocks C of σ and D of τ . When σ and τ are independent we can display the sample space Ω as a "checkerboard"



whose rows are blocks of τ, whose columns blocks of σ, and such that the "area" is proportional to the probability.

The meet is the analog for partitions of the intersection of sets. There is a whole algebra of partitions analogous to that for sets. For example, there is an analog of set union called the join of partitions and written $\sigma \vee \tau$. We leave it as an exercise to decide how this ought to be defined. We will not have need of this particular operation.

Another notion from sets is that of subset, and its analog for partitions will be very important for us. We say that a partition σ with blocks $C_1, C_2, \ldots, C_\ell$ is finer than a partition τ with blocks $D_1, D_2, \ldots, D_m$ if every block C_i of σ is contained in some block D_j of τ. We write $\sigma \leq \tau$ for this relation. If X and Y are finite-valued random variables, then $\pi(X) \leq \pi(Y)$ means that an observation of X is sufficient to determine anything one might ask about Y. The technical term for this relation is that X is a sufficient statistic for Y. More generally, if $X_1, X_2, \ldots, X_n$ are a collection of finite-valued random variables such that $\pi(X_1, \ldots, X_n)$ is finer than $\pi(Y)$, we say that $X_1, X_2, \ldots, X_n$ is sufficient for Y. In practice one often finds that in a particular experiment one wants the value of Y but that the random variables one actually measures form a sequence $X_1, X_2, \ldots$. If for some $n, X_1, X_2, \ldots, X_n$ is sufficient for Y, then one can in principle compute Y from the measurements of the X's. One also says that $X_1, \ldots, X_n$ code for Y.

Entropy

The reason for introducing partitions is that the "information content" of a finite-valued random variable X is a property of the collection of events defined by X and not by the particular labels X happens to assign to these events. We now make this precise. The entropy of a partition π whose blocks are the events $B_1, B_2, \ldots, B_n$ is defined by

$$H_2(\pi) = \sum_i P(B_i) \log_2\left(\frac{1}{P(B_i)}\right),$$

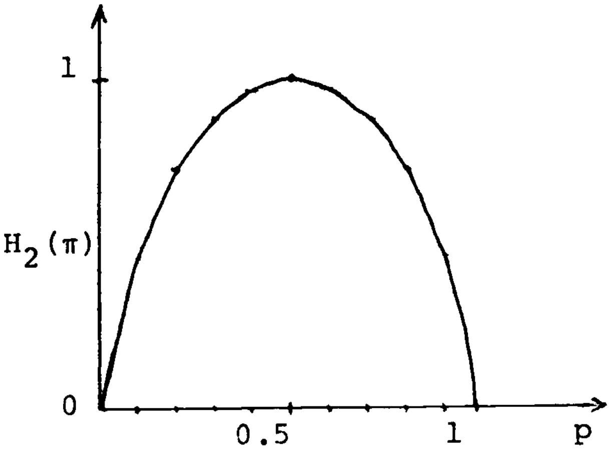
where by convention $0 \cdot \log_2(\frac{1}{0})$ is defined to be 0. The entropy of a finite-valued random variable X is the entropy of its partition:

$$H_2(X) = H_2(\pi(X)).$$

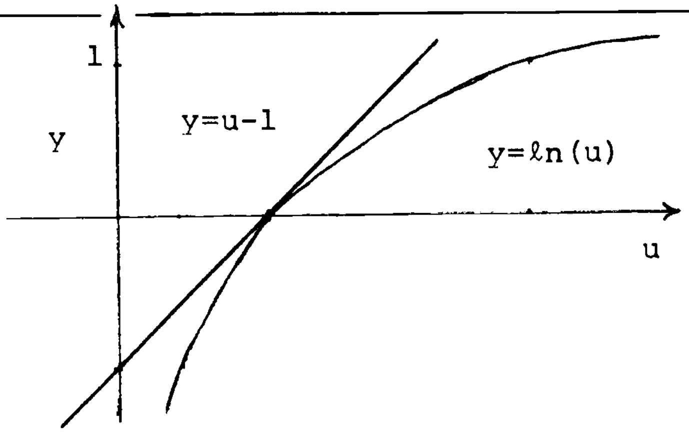
We remark that $\log_2$ could be replaced by $\log_b$ for any base $b>0$. The only effect on $H_2(\pi)$ is to multiply by the scale factor $\log_b(2)$, i.e. we merely alter the units in which the entropy is measured. The use of $\log_2$ is traditional. In this case we say that $H_2(\pi)$ is measured in <u>bits.</u> More generally, we will write $H_b(\pi)$ for $\sum_i P(B_i) \log_b(\frac{1}{P(B_i)})$. If we use $H(\pi)$ without a subscript we mean that the base b should be taken to be e, the base of the natural logarithms. We say that $H(\pi)$ is measured in <u>nats</u> (natural digits).

Consider the example of tossing a biased coin with bias p, i.e. consider a partition consisting of one or two blocks. If p is 1, then we know for certain that the coin will always show heads. In this case $H_2(X) = 0\cdot\log_2(\frac{1}{0}) + 1\cdot\log_2(\frac{1}{1}) = 0$. Entropy zero corresponds to <u>total certainty</u>. Now suppose that p is somewhat less than 1. The toss is now somewhat less predictable, and we find that the entropy is a small positive number. As p decreases, the entropy gradually increases, reaching a maximum when p = 1/2. For a fair coin

$H_2(X) = \frac{1}{2}\cdot\log_2(2) + \frac{1}{2}\cdot\log_2(2) = \frac{1}{2} + \frac{1}{2} = 1$ bit. Finally, as p decreases from 1/2 to 0, the entropy again decreases to zero; for now the toss is becoming increasingly predictable.



More generally suppose that π has n blocks. Shannon proved that $H_2(\pi)$ takes its maximum value precisely when all n outcomes are equally likely. In this case the entropy is $H_2(\pi) = \log_2(n)$ bits or $H(\pi) = \ln(n)$ nats. We will now prove this. All of our later characterizations of distributions having maximum entropy rely on the same basic technique we will use in this case. The key fact is this inequality:



$\ell n(u) \leq u-1$ for all $u > 0$

and $\ell n(u) = u-1$ if and only if $u = 1$

Basic logarithmic inequality

This fact is easy to prove using Calculus: $f(u) = \ell n(u)-u+1$ has derivative $f'(u) = \frac{1}{u} - 1$ so $f'(u)>0$ for $u<1$ and $f'(u)<0$ for $u>1$, i.e. $f(u)$ takes its maximum value at $u=1$.

Now compare $H(\pi)$ to $\ell n(n)$ using the above inequality:

$$H(\pi) - \ell n(n) = \sum_{i=1}^{n} P(B_i) \left(\ell n \left(\frac{1}{P(B_i)} \right) - \ell n(n) \right]$$

$$= \sum_{i=1}^{n} P(B_i) \, \ell n \left(\frac{1}{P(B_i)n} \right)$$

$$\leq \sum_{i=1}^{n} P(B_i) \left[\frac{1}{P(B_i)n} - 1 \right]$$

$$= \sum_{i=1}^{n} (\frac{1}{n}) - \sum_{i=1}^{n} P(B_i)$$

$$= 1 - 1 = 0.$$

The fact that the probabilities of the n blocks add up to 1, $\sum_{i=1}^{n} P(B_i) = 1$, is used twice above: in the first equality and in the second last one. In any case we find that if π has n blocks then $H(\pi) \leq \ell n(n)$.

When is equality possible? In our derivation of $H(\pi) \leq \ell n(n)$, equality can fail in only one of the steps: $\ell n\left(\frac{1}{P(B_i)n}\right) \leq \frac{1}{P(B_i)n} - 1$ for all i. Now the basic logarithmic inequality tells us that this will be an equality if and only if $\frac{1}{P(B_i)n} = 1$ for all blocks, i.e. $P(B_i) = \frac{1}{n}$ for all i. This completes our proof.

When X has a partition π all of whose blocks have the same probability, we say that X is <u>completely random</u> or <u>totally random</u>, although this is not quite the best terminology. One should really say that X has <u>maximum uncertainty</u>. (Equivalently, the measurement of X gives one the <u>maximum information</u> about the outcome of an experiment, of any random variable having the same number of outcomes.) It is unfortunate that it has become standard terminology to describe such random variables as being simply "random". For example, one often says "choose a card at random" rather than "choose a card completely at random", as if there were no other way to choose a card from a deck. In fact most "random" suffles of a deck are far from being completely random (see exercise 4); as a result, choosing a card or dealing a hand is not totally random and the probabilities computed in exercise III.4 would seldom be achieved in an actual game. On the other hand, the terminology suggests that they are. This is the price one pays for using a vague, imprecise language to describe probabilitic concepts.

Properties of Entropy

So far we have discussed examples of the entropy of some random variables. Although these examples provide some motivation for our definition of entropy they leave unanswered the basic question of why this formula and not some other is the one we use to define entropy. We will now consider why our formula is the only possible one. We will do this by finding three self-evident properties that ought to hold for any reasonable measure of information (or entropy). It then turns out that our definition of entropy is the only one that satisfies all these properties.

We begin with the most obvious of properties. As we have defined it, H is a function of partitions of the sample space. However, it should be clear that we want H to depend only on the set of probabilities of the blocks of the partition. In fact, we want H to depend only on the positive probabilities which occur. Moreover, we want H to be a continuous function of these probabilities. This is a convenience only. We could, with a great deal of effort, derive continuity from other more complex conditions; but we would rather concentrate on the important issues. We summarize the conditions on H we have just described before going on to the difficult question of conditional entropy.

Entropy property 1. An entropy is a function H defined on sets $\{p_1, p_2, \ldots, p_n\}$ of nonnegative real numbers, which satisfy $p_1+p_2+\cdots+p_n = 1$.

Entropy property 2. If H is an entropy function, then for any set $\{p_1, p_2, \ldots, p_n\}$ on which H is defined, H satisfies:

$$H(p_1, p_2, \ldots, p_n, 0) = H(p_1, p_2, \ldots, p_n).$$

In other words, H depends only on the nonzero p_i's in a given set.

Entropy property 3. An entropy function is continuous.

There are two ways to think of the concept of conditional entropy, and the fact that they are equivalent is our next property of entropy. To illustrate the ideas involved, we consider the following simple weighing problem. We have three coins, some of which may be counterfeit (but not all). Counterfeit coins are distinguishable from normal coins by the fact that they are lighter. We are given a balance scale, and we wish to find out which, if any, of the coins are counterfeit. The sample space for this problem consists of seven sample points, one for each possible set of good coins. We denote them as follows:

$$\Omega = \{1, 2, 3, 12, 13, 23, 123\}.$$

Now what happens when we put the first two coins on each side of the scale? The sample space is partitioned into three blocks corresponding to the three possible outcomes of the weighing: $\sigma = \{12, 123, 3\}$, $\{2, 23\}$, $\{1, 13\}$. After recording the result of this weighing, we then place the second and third coins on the two sides of the scale. The result of this second weighing is to partition each of the blocks of the first weighing:

$$\{12,123,3\} \text{ becomes } \{12\},\ \{123\},\ \{3\}$$
$$\{2,23\} \text{ becomes } \{2\},\ \{23\}$$
$$\{1,13\} \text{ becomes } \{1\},\ \{13\}.$$

The combined information of the two weighings is represented by the partition into seven blocks, each with one sample point. Call this partition π. Conditional entropy is concerned with the effect of the second weighing, _given_ that the first has occurred. One way to analyze this is to look at each block σ_i of the partition of the first weighing and to analyze the situation as if σ_i were the whole sample space. In general, for an event A and a partition τ we define the _conditional entropy of τ given_ A, written $H(\tau | A)$, to be the entropy of the partition $\tau_1 \cap A$, $\tau_2 \cap A, \ldots$ that τ induces on A. Thus in the above weighing problem we have three conditional entropies, one for each possible outcome of the first weighing: $H(\pi \mid \sigma_1)$, $H(\pi \mid \sigma_2)$ and $H(\pi \mid \sigma_3)$. The conditional entropy of π given σ is then defined to be the average of these. More precisely, if π and σ are any two partitions of a sample space Ω such that π is finer than σ, we define the *conditional entropy of* π *given* σ to be the average value of $H(\pi \mid \sigma_i)$ over all blocks σ_i of σ:

$$H(\pi \mid \sigma) = \sum P(\sigma_i) H(\pi \mid \sigma_i).$$

On the other hand, we would like to think of information as a "quantity" that increases as we ask more and more questions about our experiment. Therefore, the conditional entropy of π given σ ought to be the net increase in entropy from σ to π. In other words, we require our entropy function to satisfy:

Entropy property 4. If π is a finer partition than σ, then

$$H(\pi \mid \sigma) = H(\pi) - H(\sigma).$$

The last property we require is one that we have already discussed. The partition having maximum entropy among all partitions with a given number of blocks is the one for with all the blocks have the same probability.

Entropy property 5. If H is an entropy function, then for any set $\{p_1, p_2, \ldots, p_n\}$ on which H is defined, H satisfies:

$$H(p_1, p_2, \ldots, p_n) \leq H(\tfrac{1}{n}, \tfrac{1}{n}, \ldots, \tfrac{1}{n}).$$

We are now ready for the following remarkable fact: if H satisfies the above five properties, then H is given by the formula introduced earlier in this chapter, except for a possible scale change.

Uniqeness of Entropy If H is a function satisfying the five properties of an entropy function, then there is a constant C such that H is given by:

$$H(p_1,p_2,\ldots,p_n) = C\sum_i p_i \log_2(p_i).$$

Proof

The proof is rather technical, so we suggest omitting it on the first reading, returning to it later. We first apply property 4 to the partition consisting of just one block: Ω itself. By definition $H(\Omega|\Omega)$ is the same as $H(\Omega)$. Therefore, $H(\Omega) = H(\Omega) - H(\Omega) = 0$.

We now define a function f(n) by $H(\frac{1}{n}, \frac{1}{n},\ldots, \frac{1}{n})$. We have just shown that $f(1) = 0$, and we want to calculate f(n) in general. Using properties 2 and 5, we show that f(n) is increasing:

$$f(n) = H(\tfrac{1}{n},\ldots, \tfrac{1}{n}) = H(\tfrac{1}{n},\ldots,\tfrac{1}{n},0) \leq H(\tfrac{1}{n+1},\ldots, \tfrac{1}{n+1}) = f(n+1).$$

Next we consider a partition σ consisting of n^{k-1} blocks each of which has probability $\frac{1}{n^{k-1}}$. Then subdivide each of these into n parts, each of which has the same probability. call the resulting partition π. The conditional entropy $H(\pi|\sigma_i)$ for each block σ_i is clearly given by f(n). Thus the conditional entropy $H(\pi|\sigma)$ is f(n). By property 4, $f(n) = H(\pi|\sigma) = H(\pi) - H(\sigma) = f(n^k) - f(n^{k-1})$. Therefore, if we apply this fact k times, we obtain: $f(n^k) = kf(n)$.

Now fix two positive integers n and k. Since the exponential function is an increasing function, there is an integer b such that: $2^b \leq n^k \leq 2^{b+1}$. We now apply the two facts about f(n) obtained above to this relation:

$$f(2^b) \leq f(n^k) \leq f(2^{b+1}) \qquad \text{(f is increasing)}$$

$$bf(2) \leq kf(n) \leq (b+1)f(2)$$

Now divide these inequalities by kf(2):

$$\frac{b}{k} \leq \frac{f(n)}{f(2)} \leq \frac{b+1}{k}.$$

Now apply the increasing function $\log_2$ to the inequalities $2^b \leq n^k \leq 2^{b+1}$. This gives that $b \leq k \log_2(n) \leq b+1$. If we divide these by k we obtain:

$$\frac{b}{k} \leq \log_2(n) \leq \frac{b+1}{k}.$$

It follows that both f(n)/f(2) and $\log_2(n)$ are in the interval $\left[\frac{b}{k}, \frac{b}{k}+\frac{1}{k}\right]$. This implies that f(n)/f(2) and $\log_2(n)$ can be no farther apart than $\frac{1}{k}$, the length of this interval. But n and k were arbitrary positive integers. So if we let k get very large, we are forced to conclude that f(n)/f(2) coincides with $\log_2(n)$. Thus for positive integers n, we have:

$$f(n) = f(2)\log_2(n).$$

We will define the constant C to be -f(2). Since $f(2) \geq f(1) = 0$, we know that C is negative.

We next consider a set $\{p_1, p_2, \ldots, p_n\}$ of positive rational numbers such that $p_1+p_2+\cdots+p_n = 1$. Let N be their common denominator, i.e., $p_i = a_i/N$, for all i, where each a_i is an integer and $a_1+a_2+\cdots+a_n = N$. Let σ be a partition corresponding to the set of probabilities $\{p_1, p_2, \ldots, p_n\}$. Let π be a partition obtained by breaking up the i^{th} block of σ into a_i parts. Then every block of π has probability $1/N$. By definition of conditional entropy, $H(\pi | \sigma_i) = f(a_i)$ and $H(\pi | \sigma) = \sum_i p_i H(\pi | \sigma_i) = \sum_i p_i \; f(a_i) = -C \sum_i p_i \log_2(a_i)$.

By property 4, on the other hand, we have:

$$H(\pi | \sigma) = H(\pi) - H(\sigma) = f(N) - H(\sigma) = -C\log_2(N) - H(\sigma).$$

Combining the two expressions for $H(\pi | \sigma)$ gives us:

$$\begin{aligned} H(\sigma) &= -C\log_2(N) + C \sum_i p_i \log_2(a_i) \\ &= C \left[-\sum_i p_i \; \log_2(N) + \sum_i p_i \; \log(a_i) \right] \\ &= C \; \left[\sum_i p_i \; (\log_2(a_i) - \log_2(N)) \right] \\ &= C \; \left[\sum_i p_i \; \log_2(a_i/N) \right] \\ &= C \sum_i p_i \; \log(p_i). \end{aligned}$$

By continuity (property 3), H must have this same formula for all sets $\{p_1, p_2, \ldots, p_n\}$ on which it is defined. This completes the proof.

We leave it as an exercise to show that the above formula for entropy actually satisfies the five postulated properties. We conclude by giving an interpretation of independence of partitions in terms of conditional entropy. Intuitively if π and σ are independent then their joint entropy $H(\pi \wedge \sigma)$ is the sum of the individual entropies: $H(\pi) + H(\sigma)$. In terms of conditional entropy, this says that $H(\pi \wedge \sigma | \sigma) = H(\pi)$.

2*. The Shannon Coding Theorem

A consequence of Entropy property 4 of the last section is that if we wish to answer a question X by means of a sequence of questions $S_1, S_2, \ldots, S_n$, the joint entropy of $S_1, S_2, \ldots, S_n$ must be at least as large as the entropy of X, and hence the sum of the entropies of the S_i's must be at least as large as the entropy of X. In particular, if the S_i's are yes-no questions, then $H_2(S_i) \leq 1$ and we get the crude inequality $n \geq H_2(X)$. The problem of finding a set of sufficient statistics for a random variable X is called the coding problem for X, and the sequence $S_1, S_2, \ldots, S_n$ is said to code X. As we will see in the exercises, the kinds of questions one may ask are usually restricted to some class of questions. Devising particular codes is a highly nontrivial task.

One of the reasons that coding is so nontrivial in general is that one is usually required to answer a whole sequence of questions $X_1, X_2, \ldots$ produced by some process, and as a result one would like to answer the questions in the most efficient way possible. Consider one example. Suppose that X takes value 0 with probability 0.85 and takes values 1 through 200 each with probability 7.5×10^{-4}. Then $H_2(X)$ is less than 1. Simply by counting one can see that at least 8 yes-no questions will be needed to achieve a sufficient statistic for X, even though the entropy suggests that one should be able to determine X with a single yes-no question.

Shannon's Theorem states that for any finite-valued random variable X, it is possible to encode efficiently a sequence of independent copies of X provided that:

(1) one encodes a block $X_1, X_2, \ldots, X_n$ all at one time,

(2) one is willing to accept a small probability of error, $\varepsilon > 0$, that a block is incorrectly coded, such that ε can be made arbitrarily small.

Since one frequently encounters sequences of random variables in actual practice, it is not unreasonable to encode them in blocks. The small probability of error is also acceptable since it can be made arbitrarily small. Consider for example the random variable X mentioned in the preceding paragraph. Since $H_2(X) < 1$, Shannon's Theorem says that there is a block size n such that a sequence of n independent copies of X, $X_1, \ldots, X_n$,can be encoded with a sequence of n yes-no questions $S_1,\ldots,S_n$. Consider that the sequence of X_i's can take one of 201^n values, while the sequence of S_i's takes on at most 2^n possible values and you will begin to appreciate Shannon's Theorem.

We must first make precise the idea that a sequence of random variables "almost" codes for another sequence. Let $X_1,\ldots,X_n$ and $S_1,\ldots,S_r$ be two sequences of random variables. We say that $S_1,\ldots,S_r$ is *almost sufficient* for $X_1,\ldots,X_n$ with *confidence* $1-\varepsilon$ if there is an event A such that

(1) $P(A) = 1-\varepsilon$

(2) $S_1|A,\ldots,S_r|A$ is sufficient for $X_1|A,\ldots,X_n|A$, where $X_i|A$ is the random variable X_i conditioned by the occurrence of A.

Put another way, condition (2) says that the joint partition $\pi(S_1)\wedge\ldots\wedge\pi(S_r)$ when restricted to A is finer than $\pi(X_1)\wedge\ldots\wedge\pi(X_n)$ when restricted also to A.

Shannon's Coding Theorem. Let $X_1,X_2,\ldots,$ *be a sequence of independent equidistributed finite-valued random variables such that* $H_2(X_i) = h$. *For any* $\varepsilon>0$ *no matter how small and any* $\delta>0$ *no matter how small, there is an integer* N *such that for any block size* $n\geq N$, *one can find a sequence* $S_1,S_2,\ldots,S_{[hn+\delta n]}$ *of* $[hn+\delta n]$ *random variables each taking two values, which is almost sufficient for* $X_1,X_2,\ldots,X_n$ *with confidence* $1-\varepsilon$.

The confidence $1-\varepsilon$ represents the probability that the S_i's are able to code for a particular sequence of values of $X_1,X_2,\ldots,X_n$. The expression $[hn+\delta n]$ stands for the smallest integer larger than $hn+\delta n$. Finally, by entropy considerations we know that at least $[nh]$ S's will be needed to code for $X_1,X_2,\ldots,X_n$. The additional δn S's represent an extra set of S's beyond those required by entropy, but they can be chosen to be as small a fraction of the total set of S's as we please.

Proof. One begins by defining a sequence of random variables $Y_1,Y_2,\ldots$ by decreeing that if X_i takes value n then Y_i takes value $P(X_i=n)$. For example, if X_i took values $1,\ldots,n$ each with probability $\frac{1}{n}$, then Y_i would take value $\frac{1}{n}$ with probability 1.

These random variables have two properties we need. The first is that the Y_i's are independent. This is an immediate consequence of the fact that the X_i's are so. The second fact is that the expected value of $\log_2(1/Y_i)$ is h, the entropy of X_i. To see this we simply compute:

$$E(\log_2(1/Y_i)) = \sum_n \log_2(1/P(X_i=n))P(X_i=n)$$

$$= H_2(X_i) = h \;,$$

since $\log_2(1/Y_i)$ takes value $\log_2(1/P(X_i=n))$ when $X_i = n$.

The sequence $\log_2(1/Y_1)$, $\log_2(1/Y_2)$,... is a sequence of independent equidistributed random variables each with mean h. By the Law of Large Numbers,

$$P\left(\lim_{n\to\infty} \frac{\log_2(1/Y_1)+\log_2(1/Y_2)+\ldots+\log_2(1/Y_n)}{n} = h\right) = 1.$$

Now $\log_2(1/Y_1)+\log_2(1/Y_2)+\ldots+\log_2(1/Y_n) = \log_2\left(\frac{1}{Y_1Y_2\ldots Y_n}\right)$. Therefore:

$$P\left(\lim_{n\to\infty} \frac{1}{n}\log_2(\frac{1}{Y_1Y_2\ldots Y_n}) = h\right) = 1.$$

This says that for n large enough the expression $\frac{1}{n}\log_2\left(\frac{1}{Y_1Y_2\ldots Y_n}\right)$ will be as close to h as we please with as high a probability as we please. The probability we want is 1-ε, and we want $\frac{1}{n}\log_2\left(\frac{1}{Y_1Y_2\ldots Y_n}\right)$ to be within δ of h with this probability:

$$P\left(\left|\frac{1}{n}\log_2\left(\frac{1}{Y_1Y_2\ldots Y_n}\right) - h\right| < \delta\right) = 1-\varepsilon.$$

As one might expect, the event A in the definition of a set of almost sufficient statistics will be the above event:

$$A = \left(\left|\frac{1}{n}\log_2\left(\frac{1}{Y_1Y_2\ldots Y_n}\right) - h\right| < \delta\right).$$

$$= (\left|\log_2\left(\frac{1}{Y_1Y_2\ldots Y_n}\right) - nh\right| < n\delta)$$

$$= (-n\delta < \log_2\left(\frac{1}{Y_1Y_2\ldots Y_n}\right) - nh < n\delta).$$

Exponentiating every term in the above pair of inequalities preserves the inequalities so

$$A = (2^{-n\delta} < \frac{1}{Y_1 Y_2 \ldots Y_n} \cdot 2^{-nh} < 2^{n\delta})$$

$$= (2^{-n\delta+nh} < \frac{1}{Y_1 Y_2 \ldots Y_n} < 2^{n\delta+nh})$$

$$= (2^{-nh+n\delta} > Y_1 Y_2 \ldots Y_n > 2^{-nh-n\delta}).$$

We are now ready for the crucial step in the proof. We count how many blocks of the joint partition $\pi(X_1) \wedge \pi(X_2) \wedge \ldots \wedge \pi(X_n)$ are contained in the event A. Suppose that there are r such blocks; call them $B_1, B_2, \ldots, B_r$. Each of these blocks is of the form $(X_1 = i_1) \cap (X_2 = i_2) \cap \ldots \cap (X_n = i_n)$. If we sum the probabilities of <u>all</u> such events we get 1:

$$\sum_{i_1, \ldots, i_n} P((X_1 = i_1) \cap (X_2 = i_2) \cap \ldots \cap (X_n = i_n)) = 1.$$

Since the X_i's were assumed to be independent, this means that

$$\sum_{i_1, \ldots, i_n} P(X_1 = i_1) P(X_2 = i_2) \ldots P(X_n = i_n) = 1.$$

Now each of the above factors is, by definition, the value that the corresponding Y_i takes:

$$\sum_{\substack{X_1=i_1\\X_2=i_2\\\cdots\\X_n=i_n}} Y_1\cdot Y_2\cdots Y_n = 1 \;.$$

If we sum just over those blocks contained in A we find that

$$\sum_{j=1}^{r} Y_1\cdot Y_2\cdots Y_n \leq 1 \;,$$

each Y_i taking the value appropriate to the blocks B_j. But for these blocks we know that $Y_1Y_2\ldots Y_n>2^{-nh-n\delta}$. Hence

$$\sum_{j=1}^{r} 2^{-nh-n\delta} \leqslant \sum_{j=1}^{r} Y_1\cdot Y_2\cdots Y_n \leq 1.$$

Now the terms of the sum $\sum_{j=1}^{r} 2^{-nh-n\delta}$ do not depend on the block B_j. So we find that

$$r\cdot 2^{-nh-n\delta} < 1$$

or that $\qquad r<2^{nh+n\delta}$,

i.e. there are fewer than $2^{nh+n\delta}$ blocks in A.

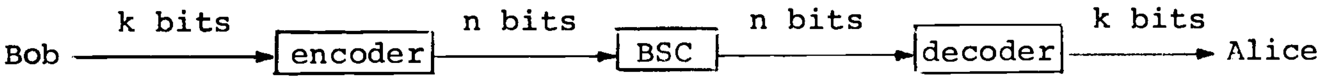
We are now ready to code the random variables $X_1, X_2, \ldots, X_n$. Number the blocks in A in binary using $[nh+n\delta]$ binary digits starting with 00...01 and ending with 11...11. Note that because $r<2^{nh+n\delta}$, we will have at most $2^{[nh+n\delta]}-1$ blocks in A. We assign the binary number 00...00 to all blocks outside A.

The random variable S_i is defined to be the i^{th} digit of the block in which the outcome occurs. By definition $S_1, S_2, \ldots, S_{[nh+n\delta]}$, when restricted to A, are sufficient for $X_1, X_2, \ldots, X_n$ when restricted to A. When all the S_i's take the value 0, we are unable to determine the values of $X_1, X_2, \ldots, X_n$, but when the S_i's take any other set of values, we can compute all the values of $X_1, X_2, \ldots, X_n$.

This completes the proof.

The usual form in which one sees this theorem is called the Shannon Channel Coding Theorem. The problem here is to transmit information through a noisy channel. The channel we consider is called the Binary Symmetric Channel. Each bit of information one transmits through the BSC is either left alone or changed. The probability that it is changed is p, the same for all bits, and each bit is altered or not independently of the others. The BSC is equivalent to the Bernoulli process, coin tossing, with bias p.

Transmission through the BSC proceeds as follows. A message k bits long is first sent through an <u>encoder</u> where it is changed into a string of n bits. This string of bits is then transmitted through the BSC to a <u>decoder</u> that converts the received n bits into a string of k bits, which we hope is the same as the original message. The problem is to design the encoder and decoder so that the probability of error per transmitted bit is smaller than some preassigned value and so that the <u>redundancy</u> $\frac{n}{k}$ is as small as possible. Equivalently, we want the <u>rate of transmission</u> $\frac{k}{n}$ to be as high as possible.



We may think of the noise as a sequence of Bernoulli random variables $X_1, X_2, \ldots, X_n$ that are added to the signal. Let $h = H_2(X_i) = p \log_2(\frac{1}{p}) + q \log_2(\frac{1}{q})$. Then the input signal plus the noise constitute a total entropy of k+nh bits. The decoder can ask at most n questions about the data it receives, since it receives just n bits of data. From these n questions it must determine both the noise and the original signal, hence $k+nh \leq n$. Put another way, the decoder may ask just n-k questions in order to determine the noise and eliminate it. Thus $k+nh \leq n$ or $n-nh \geq k$. Hence $1-h \geq k/n$.

This says that the rate of communication through the BSC can never be greater than 1-h. One calls 1-h the capacity of the channel.

The Shannon Channel Coding Theorem says that for any rate r less than the channel capacity 1-h it is possible to choose k and n so that k/n>r and to design an encoder and decoder so that the (average) probability of error per message bit is as small as we please. The proof is very similar to the proof we just gave for the Shannon Coding Theorem.

3. Continuous Entropy

We now consider what entropy means for continuous random variables. The concepts in this case are by no means as self-evident as in the case of a finite-valued random variable.

Relative Entropy

The most obvious way to begin is to try to "finitize". Let X be a continuous random variable taking values in some finite interval. For simplicity take this interval to be [0,a]. Now exactly as in Calculus, we partition (or subdivide) this interval into n blocks $B_1, B_2, \ldots, B_n$. The i^{th} block is the subinterval [(i-1)a/n, ia/n). Define a new random variable Y_n that takes value (i-1)a/n whenever X takes a value in the block B_i. We call Y_n the n^{th} truncation of X. We show why we use this name by means of an example. Suppose that a=1 and that n = 1000. Now imagine that we perform our experiment and that the outcome X is

$$.1415926\ldots\ .$$

The value of Y_{1000} in this case would be

$$.141000\ldots,$$

i.e. we truncate the value of X to 3 decimal places. Clearly the truncations of X will be better and better approximations to X as $n \to \infty$. Moreover the truncations are finite-valued random variables.

One might add that in practice one always uses a truncation in an actual experiment. It is only in our idealized mathematical models that one can speak of an arbitrary real number.

Now compute the entropy of Y_n. By definition of Y_n,

$$P\left\{Y_n = \frac{(i-1)a}{n}\right\} = P\left\{\frac{(i-1)a}{n} \le X < \frac{ia}{n}\right\} = F\left(\frac{ia}{n}\right) - F\left(\frac{(i-1)a}{n}\right),$$

where F(t) is the probability distribution of X. Thus

$$H(Y_n) = \sum_{i=1}^{n} P\left\{Y_n = \frac{(i-1)a}{n}\right\} \ell n\left(\frac{1}{P\left\{Y_n = \frac{(i-1)a}{n}\right\}}\right)$$

$$= \sum_{i=1}^{n} \left\{F\left(\frac{ia}{n}\right) - F\left(\frac{(i-1)a}{n}\right)\right\} \ell n\left\{\frac{1}{F\left(\frac{ia}{n}\right) - F\left(\frac{(i-1)a}{n}\right)}\right\}.$$

The crucial step in the computation is the mean value theorem of Calculus: if F is differentiable on the interval [s,t], then for some x between s and t

$$F'(x)(t-s) = F(t)-F(s).$$

We apply this to each block B_i. Each block has length $\frac{a}{n}$, so

$$H(Y_n) = \sum_{i=1}^{n} F'(x_i)\ \frac{a}{n}\ \ell n\left(\frac{1}{F'(x_i)\, a/n}\right)$$

$$= \sum_{i=1}^{n} F'(x_i)\ \frac{a}{n}\ \ell n\ \left(\frac{1}{F'(x_i)}\right) + \sum_{i=1}^{n} F'(x_i)\frac{a}{n}\ \ell n\left(\frac{n}{a}\right)\ ,$$

where each x_i is some point in the block B_i. If we write $f(t) = F'(t)$ for the density of X, then the first term above is

$$\sum_{i=1}^{n} f(x_i)\ \ell n\left(\frac{1}{f(x_i)}\right)\ \frac{a}{n}\ .$$

This is just the Riemann sum for our partition of [0,a]. So as n→∞ this approaches

$$\int_0^a f(x)\ \ell n\left(\frac{1}{f(x)}\right)\ dx.$$

Next consider the second term above. We may write this as

$$\ell n\left(\frac{n}{a}\right)\ \sum_{i=1}^{n} f(x_i)\frac{a}{n}\ .$$

Except for the factor $\ell n(\frac{n}{a})$, we would have a Riemann sum for $\int_0^a f(x)\,dx = 1$. However the factor $\ell n(\frac{n}{a})$ means that as n→∞,

$$H(Y_n) \approx \int_0^a f(x)\ \ell n\left(\frac{1}{f(x)}\right) dx + \ell n(n) - \ell n(a).$$

So as $n\to\infty$, $H(Y_n)\to\infty$.

The difficulty is easily seen. As we partition $[0,a]$ into finer and finer blocks, the random variable Y_n is taking an enormous number of values, some fraction of which are roughly equally likely. This is an artifact of our subdivision process and ought to be eliminated. We do this by measuring not the absolute entropy of Y_n but rather the difference between the entropy of Y_n and the maximum possible entropy of a random variable taking n values. We call this the <u>relative entropy</u> of Y_n:

$$\text{Relative entropy of } Y_n = H(Y_n) - \ln(n).$$

In other words, instead of measuring how far Y_n is from being completely certain, we measure how close Y_n is to being completely random. For finite-valued random variables these two ways of measuring entropy are equivalent, but when we take the limit as $n\to\infty$, only the relative entropy converges. We therefore define:

$$\begin{aligned}\text{Relative entropy of } X &= \lim_{n\to\infty} (\text{relative entropy of } Y_n) \\ &= \int_0^a f(x)\ln\left(\frac{1}{f(x)}\right)dx - \ln(a).\end{aligned}$$

For this notion of entropy, the case of total randomness will be represented by a relative entropy of zero. Less uncertain random variables will have a negative relative entropy. Continuous random variables can have arbitrarily large negative entropy: complete certainty is impossible for continuous random variables.

Which continuous random variables will have entropy zero? In other words, what is the continuous analogue of the equally likely probability distribution? To answer this we proceed as we did for finite-valued random variables. Let X be any continuous R.V. taking values in [0,a]. Then

$$\text{relative entropy of } X = \int_0^a f(x)\,\ell n\left(\frac{1}{f(x)}\right)dx - \ell n(a)$$

$$= \int_0^a f(x)\,\ell n\left(\frac{1}{f(x)}\right)dx - \ell n(a)\int_0^a f(x)\,dx$$

$$= \int_0^a f(x)\left[\ell n\left(\frac{1}{f(x)}\right) - \ell n(a)\right]dx$$

$$= \int_0^a f(x)\,\ell n\left(\frac{1}{f(x)\,a}\right)dx$$

$$\leq \int_0^a f(x)\left[\frac{1}{f(x)\,a} - 1\right]dx$$

$$= \int_0^a \frac{dx}{a} - \int_0^a f(x)\,dx$$

$$= 1 - 1 = 0 .$$

Now the logarithmic inequality tells us that the above inequality is an equality if and only if $\frac{1}{f(x)a} = 1$ or $f(x) = \frac{1}{a}$. In other words, the maximum entropy occurs precisely when X has the uniform distribution on [0,a].

Boltzmann Entropy

The notion of relative entropy is fine for random variables taking values in a finite interval, but most continuous random variables we have seen do not have this property. The most natural way to try to extend entropy to arbitrary continuous random variables is to use a limiting process similar to what we used for extending entropy from finite-valued random variables to finite-interval random variables. We will do this first for positive random variables before going on to the general case.

Let T be a positive continuous random variable (i.e. $P(T\leq 0) = 0$). For $a>0$, we define the restriction of T to [0,a] to be the random variable $T_a = T|(T\leq a)$. By this we mean that T_a takes the value of T conditioned on the occurrence of $(T\leq a)$. We already saw this in the definition of almost sufficient statistics. The probability distribution of T_a is given by $P(T_a\leq t) = P(T\leq t|T\leq a)$, and the density of T_a is then given by

$$\text{dens}(T_a=t) = \begin{cases} f(t)/\int_0^a f(u)\,du, & \text{if } 0\le t\le a \\ 0 & \text{, if } t>a \text{ or } t<0, \end{cases}$$

where $f(t) = \text{dens}(T=t)$. Write $C_a = \dfrac{1}{\int_0^a f(u)\,du}$ for the above normalization constant. Clearly T_a will be a better and better approximation to T as $a\to\infty$. As with truncations, restrictions are always used in an actual experiment.

It would be nice if we could define the relative entropy of T to be the limit of the relative entropy of T_a as $a\to\infty$, but unfortunately this diverges:

$$\text{relative entropy of } T_a = \int_0^a C_a\ f(t)\,\ell n\left(\frac{1}{C_a f(t)}\right) dt - \ell n(a) \to \infty$$

$$\text{as} \quad a\to\infty.$$

As before the difficulty is that we are not measuring entropy properly. The case of total randomness, entropy zero, is the uniform distribution on $[0,a]$;but as $a\to\infty$ this distribution ceases to make sense. So we are attempting to measure the entropy of T relative to that of a nonexistent distribution! What should we do? We no longer have either total certainty or total uncertainty from which to measure entropy.

What we do is to "renormalize" our measurement of entropy so that the entropy of the uniform distribution on [0,a] is $\ell n(a)$ rather than 0. We do this by analogy with the "equally likely" distribution on n points whose entropy is $\ell n(n)$. There is no really convincing justification for this choice of normalization. The entropy defined in this way is called the <u>Boltzmann</u> or <u>differential</u> <u>entropy</u>:

$$\begin{aligned} H(T) &= \lim_{a\to\infty}[(\text{relative entropy of } T_a) + \ell n(a)] \\ &= \lim_{a\to\infty} \int_0^a C_a \; f(t)\,\ell n\left[\frac{1}{C_a f(t)}\right] dt \\ &= \int_0^\infty f(t)\,\ell n\left[\frac{1}{f(t)}\right] dt, \end{aligned}$$

if this improper integral exists. The same definition works, in fact, for any continuous random variable.

We now ask which positive continuous random variables take maximum Boltzmann entropy. Let T be such a R.V., and let $\mu = E(T)$ be its expectation. To bound the entropy of T we use a method known as the <u>Lagrange</u> <u>multiplier</u> <u>method</u>. This method is appropriate wherever we wish to maximize some quantity subject to constraints. In this case the constraints that the density function f(t) of T must satisfy are:

$$\int_0^\infty f(t)\,dt = 1 \quad \text{and} \quad \int_0^\infty tf(t)\,dt = \mu.$$

Multiply the constraints by constants α and β to be determined later and subtract both from the entropy of T. Then proceed as in all our previous maximum entropy calculations:

$$H(T) -\alpha-\beta\mu = \int_0^\infty f(t)\,\ell n\left(\frac{1}{f(t)}\right)dt - \alpha\int_0^\infty f(t)\,dt-\beta\int_0^\infty tf(t)\,dt$$

$$= \int_0^\infty f(t)\left[\ell n\left(\frac{1}{f(t)}\right) -\alpha-\beta t\right]dt$$

$$= \int_0^\infty f(t)\,\ell n\left(\frac{1}{f(t)e^{\alpha+\beta t}}\right)dt$$

$$\leq \int_0^\infty f(t)\left(\frac{1}{f(t)e^{\alpha+\beta t}} - 1\right)dt$$

$$= \int_0^\infty e^{-\alpha-\beta t}dt - \int_0^\infty f(t)\,dt$$

$$= \left[\frac{e^{-\alpha-\beta t}}{\beta}\right]_0^\infty - 1$$

$$= \frac{e^{-\alpha}}{\beta} - 1 \qquad (\text{if } \beta>0).$$

By the basic logarithmic inequality, the above inequality is an equality if and only if $\frac{1}{f(t)e^{\alpha+\beta t}} = 1$ or $f(t) = e^{-\alpha-\beta t}$.

We now use the constraints to solve for α and β:

$$1 = \int_0^\infty f(t)dt = \int_0^\infty e^{-\alpha-\beta t}dt = \frac{e^{-\alpha}}{\beta}$$

$$\mu = \int_0^\infty tf(t)dt = \int_0^\infty te^{-\alpha-\beta t}dt = \frac{e^{-\alpha}}{\beta^2} .$$

Therefore $\beta = e^{-\alpha} = \beta^2\mu$ or $\beta = 1/\mu = e^{-\alpha}$. The function f(t) thus has the form

$$f(t) = \frac{1}{\mu} e^{-t/\mu},$$

i.e. T is exponentially distributed with parameter $1/\mu$. Moreover, the entropy of T is $H(T) = \alpha+\beta\mu = 1+\ell n(\mu)$. Therefore we see that as μ gets large T can have arbitrarily high entropy. Thus there is no positive random variable having maximum entropy among all such random variables.

Standard Entropy

The reason we had to specify the expectation of a positive random variable in order to find the one having maximum entropy arises from an important distinction between finite entropy and Boltzmann entropy: the choice of units in which we measure our random variable alters the Boltzmann entropy but has no effect on the finite entropy. Indeed, the entropy of a finite-valued random variable depends only on the partition it defines. For example, if X is uniformly distributed on [0,1], then Y=2X is uniformly distributed on [0,2]. Although Y represents the same phenomenon as X, the difference being the units with which we measure distance, obviously an observation of X is more certain than an observation of Y (one bit more certain to be precise). More generally, for any continuous random variable $X, H(C\ X) = H(X) + \ln(C)$.

In order to speak of the entropy of the phenomenon represented by a random variable, independent of scale changes, we introduce yet one more notion of entropy. The standard entropy of a random variable is the Boltzmann entropy of its standardization. Using the notion of standard entropy we can ask an important question. Which continuous random variables have the maximum standard entropy? The answer is that, up to changes of scale, there is exactly one such random variable and it is a random variable we have not yet seen before: the normal distribution. This random variable forms the basis of the Wiener process, the last of the four principal stochastic processes of probability theory.

We now compute this random variable. Since we want the random variable to have maximum standard entropy, we may assume it is standard. Let X be such a random variable, and let f(x) be its density. We maximize H(X) by the method of Lagrange multipliers we used above, but now there are three constraints:

$$\int_{\infty}^{\infty} f(x)\,dx = 1, \quad \int_{-\infty}^{\infty} xf(x)\,dx = 0, \quad \int_{-\infty}^{\infty} x^2 f(x)\,dx = 1.$$

One of the constraints being zero we need just two parameters α and β to be determined later:

$$\begin{aligned}
H(X) - \alpha - \beta &= \int_{-\infty}^{\infty} f(x)\,\ell n\left(\frac{1}{f(x)}\right) dx - \alpha \int_{-\infty}^{\infty} f(x)\,dx - \beta \int_{-\infty}^{\infty} x^2 f(x)\,dx \\
&= \int_{-\infty}^{\infty} f(x) \left(\ell n\left(\frac{1}{f(x)}\right) - \alpha - \beta x^2\right) dx \\
&= \int_{-\infty}^{\infty} f(x)\,\ell n\left(\frac{1}{f(x)e^{\alpha+\beta x^2}}\right) dx \\
&\leq \int_{-\infty}^{\infty} f(x) \left(\frac{1}{f(x)e^{\alpha+\beta x^2}} - 1\right) dx \\
&= \int_{-\infty}^{\infty} e^{-\alpha-\beta x^2}\,dx - \int_{-\infty}^{\infty} f(x)\,dx \\
&= \frac{e^{-\alpha}}{\sqrt{\beta}} \int_{-\infty}^{\infty} e^{-u^2}\,du - 1 \qquad (u = \sqrt{\beta}\;x) \\
&= \frac{e^{-\alpha}\sqrt{\pi}}{\sqrt{\beta}} - 1.
\end{aligned}$$

The fact that $\int_{-\infty}^{\infty} e^{-u^2}\, du = \sqrt{\pi}$ is a standard fact from Calculus. The proof proceeds as follows. Let $A = \int_{-\infty}^{\infty} e^{-u^2} du$. Then since u is just a dummy variable, $A = \int_{-\infty}^{\infty} e^{-v^2} dv$ as well. Hence

$$A^2 = \int_{-\infty}^{\infty} e^{-u^2} du \int_{-\infty}^{\infty} e^{-v^2}\, dv$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-u^2-v^2} du\; dv.$$

Now switch to polar coordinates. Then $r^2 = u^2 + v^2$, $dudv = rdrd\theta$ and the limits of integration are $0 \leq \theta < 2\pi$ and $0 < r < \infty$:

$$A^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} rdrd\theta$$

$$= \int_0^{2\pi} \left[-\frac{1}{2}\, e^{-r^2} \right]_0^{\infty} d\theta$$

$$= \int_0^{2\pi} \frac{1}{2}\, d\theta = \pi.$$

Hence $A = \sqrt{\pi}$.

Returning to our bound on the entropy of H(X), we know by the basic logarithmic inequality that this will be an equality if and only if $f(x) = e^{-\alpha-\beta x^2}$. We now use the constraints to solve for α and β.

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} e^{-\alpha-\beta x^2}\,dx$$

$$= e^{-\alpha}\sqrt{\pi/\beta}$$

$$1 = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 e^{-\alpha-\beta x^2}\,dx$$

$$= \left[-\frac{xe^{-\alpha-\beta x^2}}{2\beta}\right]_{-\infty}^{\infty} + \frac{1}{2\beta}\int_{-\infty}^{\infty} e^{-\alpha-\beta x^2}\,dx$$

$$= 0 + \frac{1}{2\beta}\, e^{-\alpha}\sqrt{\pi/\beta}.$$

Therefore $e^{-\alpha} = \sqrt{\beta/\pi} = 2\beta\sqrt{\beta/\pi}$ from which we conclude that $\beta = 1/2$ and $e^{-\alpha} = 1/\sqrt{2\pi}$. Hence the maximum entropy among all standard continuous random variables is achieved precisely when

$$\text{dens}(X = x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} .$$

We say that X has the <u>standard</u> <u>normal</u> <u>distribution</u> in this case. More generally by changing the origin (zero point) and unit of measurement (scale) we get a collection of random variables, each determined by its mean and variance.

Definition. A continuous random variable X is said to have the normal or Gaussian distribution with mean m and variance σ^2 if

$$\text{dens}(X=x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} .$$

For brevity we will write that X is $N(m,\sigma^2)$. Some authors write $N(m,\sigma)$ instead of $N(m,\sigma^2)$; one should beware. We leave it as an exercise to verify that the above density really does define a probability distribution with mean m and variance σ^2 and that all of them have the standard normal distribution as their common standardization.

There are more distributions determined by maximum entropy, especially those in statistical thermodynamics and quantum mechanics, but I trust that you now see the basic ideas.

Summary

The four principal processes of probability theory are all determined by maximum entropy properties. We summarize this here.

Type of Entropy	Class of Random Variables	Definition
(Finite partition) entropy	Finite-valued	$H(\pi) = \sum_{i=1}^{n} P(B_i)\,\ell n\left(\frac{1}{P(B_i)}\right)$ $H(X) = H(\pi(X))$
Relative entropy	Continuous with values in [0,a]	$\int_0^a f(x)\,\ell n\left(\frac{1}{f(x)}\right)dx - \ell n(a)$
Boltzmann entropy	Continuous	$H(X) = \int_0^\infty f(x)\,\ell n\left(\frac{1}{f(x)}\right)dx$
Standard entropy	Continuous	$H\left(\frac{X-m}{\sigma}\right) = H(X) - \ell n(\sigma)$

Types of Entropy

Process	Distribution/Model	Class of Random Variables for which entropy is maximized
Sampling	Finite uniform on n points Placements of 1 ball into n boxes	Random variables taking at most n values
Uniform	Uniform on [0,a] Sampling one point completely at random from [0,a]	Continuous random variables taking values in [0,a] (Relative or Boltzmann entropy)
Poisson	Exponential, intensity α Continuous memoryless waiting time, intensity α	Positive continuous random variables of mean $1/\alpha$. (Boltzmann entropy)
Wiener	Normal, $N(m,\sigma^2)$ Position of a continuous random walk of rate σ^2 starting at m	Continuous random variables having mean m and variance σ^2 (Boltzmann entropy)

Maximum Entropy Distributions

Distribution	Entropy	Relative entropy	Boltzmann entropy	Standard entropy
Finite uniform on n points	$\log_2(n)$ bits	—	—	—
Uniform on [0,a]	-	0	$\ell n(a)$	$\frac{1}{2}\ \ell n(12)$ $\simeq$ 1.2425 nats
Exponential, intensity α	-	-	$1-\ell n(\alpha)$	1
Normal, $N(m,\sigma^2)$	-	-	$\ell n(\sigma\sqrt{2\pi e})$	$\ell n(\sqrt{2\pi e})$ $\simeq$ 1.4189 nats

Values of entropy

4. Exercises for

Chapter VII Entropy and Information

1. A visitor to an imaginary country finds that the inhabitants of city A always tell the truth, while the inhabitants of city B always lie. The visitor wants to know which city he is in. He may ask only yes-no direct questions. (A question such as "If I were to ask you what city I am in, what would you say?" are indirect and would only confuse an inhabitant.) How many questions must the visitor ask? Note that an inhabitant of city A could be temporarily residing in city B and vice versa.

2. You are given twelve coins, one of which is counterfeit, and a balance. The counterfeit coin is either light or heavy, you do not know which. How many weighings are necessary to determine which coin is counterfeit?

3. You are given five coins, some of which may be counterfeit. A counterfeit coin is lighter than a good coin, and all counterfeit coins weigh the same. Again you are given a balance. How many weighings are necessary to find all of the counterfeit coins?

4. A deck of 52 cards is said to have been randomly shuffled if all 52! permutations are equally likely. What we normally regard as being a random shuffle is in fact very far from random. For example, the cut-and-interlace shuffle (also called the perfect shuffle) has the following property. If a new deck (in the standard Bridge order: 2 of clubs, 3 of clubs,...,ace of spades) is perfectly shuffled, "cut" at a point 4m cards from the top, and dealt as in a Bridge game, then each of the four players will receive all the cards of one suit. It is known that frequent Bridge players are capable of consistently achieving a perfect shuffle.

   How much information is contained in a random shuffle? in a perfect shuffle? in a random cut? How many independent random cuts are needed to achieve a completely random shuffle?

5. Let N be an integer between 1 and 2000. Divide N by 6, 10, 22 and 35, and find the remainders. How much information about N do these four remainders tell you?

6. Show that the entropy of the normal distribution $N(m,\sigma^2)$ is $\log_2(\sigma\sqrt{2\pi\sigma})$ bits. Use this to answer the following question. A coin is tossed 1000 times, getting 368 heads and 632 tails. How much additional information will one more toss of the coin give one?

7. Suppose that the imaginary country of exercise 1 has another city C where the inhabitants alternately tell the truth and lie. What is the smallest number of questions the visitor must ask to find out which city he is in.

8.* Generalize problems 2 and 3 above to an arbitrary number of coins.

9.* Let f(x) be a differentiable function defined on [o,a]. Assume that f(o) = 0 and that $|f'(x)| \leq b$ for all x in [o,a]. Find an upper bound on the amount of information necessary to determine the value of f(x) at <u>every</u> $x \in [o,a]$ with an error not exceeding $\varepsilon > 0$.

10.* You are playing a variation of "20 questions." A chooses a number between 1 and 1,000,000, and B must find this number by asking yes-no questions about it, except that B asks <u>random</u> questions. How long does it take for B to find the number? Let T be the time needed for B to find the number.

11.* Let $k_1, k_2, \ldots$ be a sequence of numbers such that $\lim_{n\to\infty} (k_n - \log_2(n)) = c$. Let T_n be as in exercise 9 above but for the problem of guessing a number between 1 and 2^n. Prove that $\lim_{n\to\infty} P(T_n = k_n) = e^{-1/2c}$.

VIII Markov Chains

All the processes we have considered so far have been based on sequences of independent equidistributed random variables. We now consider processes which are based on sequences of dependent random variables but for which the dependence is of the simplest possible kind: the future depends on the present but not on the past.

1. The Markov Property

Let $X_0, X_1, X_2, \ldots.$ be a sequence of integer random variables. We think of the values of the X_n's as being the states of the Markov chain. Thus if $(X_n = i)$, we say the process is in state i at time n. Moreover, if $(X_n = i)$ and $(X_{n+1} = j)$, then we say there was a transition from state i to state j at time n.

Definition. A sequence $X_0, X_1, \ldots.$ of integer random variables forms a Markov chain if for any integers $i_0, i_1, \ldots., i_n$,

$$P(X_n = i_n | (X_0 = i_0) \cap (X_1 = i_1) \cap \cdots \cap (X_{n-1} = i_{n-1}))$$
$$= P(X_n = i_n | X_{n-1} = i_{n-1}).$$

In other words, the future states of the Markov chain are dependent only on the present state and not on how the Markov chain reached the present state. We call this condition the Markov property.

The conditional probability

$$p_{ijn} = P(X_{n+1} = j \mid X_n = i)$$

is called the <u>transition probability from state</u> i <u>to state</u> j <u>at time</u> n. By the law of alternatives, the probability distribution of X_{n+1} is determined by the transition probabilities and the probability distribution of X_n:

$$\begin{aligned} P(X_{n+1} = j) &= \sum_i P(X_{n+1} = j \mid X_n = i) P(X_n = i) \\ &= \sum_i p_{ijn} P(X_n = i). \end{aligned}$$

As a result we see that all the probability distributions of the X_n's as well as all their joint distributions are determined by the distribution of X_0 and the transition probabilities.

We have seen several examples of Markov chains already. The Bernoulli process is a Markov chain having two states: heads and tails, or 1 and 0. In this case the transition probabilities are given by

$$\begin{array}{ll} p_{00n} = q & p_{10n} = p \\ p_{10n} = q & p_{11n} = p \end{array}$$

(the probability distribution of X_0 can be anything). Another example is the sequential sampling process. Here the state is the number of red balls in the urn. For if we know the number of red balls in the urn as well as the number of balls chosen so far (i.e. the time) we can compute how many black balls are in the urn.

The sequential sampling process has the property that the transition probabilities depend not only on the states i and j but also on n, the number of balls chosen so far. In such a case our process is continually changing or inhomogeneous. In this chapter we will only study Markov chains such that the transition probabilities are independent of time.

Definition A Markov chain $X_0, X_1, \ldots$ is said to be homogeneous if the transition probabilities

$$p_{ij} = P(X_{n+1} = i \,|\, X_n = j)$$

do not depend on n.

Many apparently inhomogeneous Markov chains can be reinterpreted as homogeneous Markov chains, so that this concept is not as special as it may at first appear. For example, if we define a "state" of the sequential sampling process to be the pair of numbers: (no. of red balls, no. of black balls), then the sequential sampling process is a homogeneous Markov chain.

When we write the transition probabilities p_{ij} as a matrix we get a matrix M called the transition probability matrix of the Markov chain.

$$\text{input state} \longrightarrow \underset{\substack{\Big\downarrow \\ \text{output state}}}{\begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & p_{11} & p_{12} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & p_{21} & p_{22} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}}$$

The rows represent the starting states and the columns represent the ending states, during each unit of time. The transition probability matrix determines the Markov chain except for the probability distribution of X_0. The entries of the matrix must be between 0 and 1, and the sum of the entries of each row is 1. On the other hand, we can say nothing about the columns.

Definition A row vector (with possibly infinitely many coefficients) is said to be a stochastic vector if all entries are between 0 and 1 and the sum of all coefficients is 1. A square matrix (with possibly infinitely many rows) is called a stochastic matrix if its rows are all stochastic vectors.

The term "stochastic vector" is simply another way of looking at the probability distribution of an integer random variable. We see that distributions and Markov chains give rise to a new way of looking at vectors and matrices. A pair consisting of a stochastic matrix M and a stochastic vector $\vec{u}$ determines a unique Markov chain such that $\vec{u}$ is the row vector corresponding to the distribution of X_0 and M is the transition probability matrix.

We call the distribution of X_0 the <u>initial distribution</u> of the Markov chain. As we have already remarked, the distributions of X_1, X_2, X_3,..... are determined successively by the formula

$$P(X_{n+1} = j) = \Sigma_i \, p_{ij} P(X_n = i).$$

In terms of matrices, this says that if $\vec{u}_n$ is the stochastic vector corresponding to the distribution of X_n, then

$$\vec{u}_{n+1} = \vec{u}_n \cdot M$$

is the stochastic vector corresponding to X_{n+1}, where $\vec{u}_n \cdot M$ is the product of the matrices $\vec{u}_n$ and M. In other words, the transition from time n to time $n+1$ in a Markov chain corresponds to matrix multiplication. More generally, we can iterate the above formula to get

$$\vec{u}_n = \vec{u}_0 \cdot M^n ,$$

showing explicitly how the distributions of the X_n's depend on $\vec{u}_0$ and M.

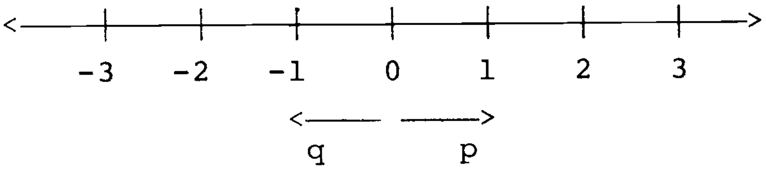
The Bernoulli Process

The Bernoulli process, as the process of tossing a coin, is a Markov chain whose transition matrix is

$$M = \begin{bmatrix} q & p \\ q & p \end{bmatrix}$$

Notice that $M^n = M$ for all n and that $\vec{u}_0 M^n = [q,p]$ no matter what the initial distribution is.

On the other hand if we use the random walk interpretation of the Bernoulli process, we get a very different Markov chain. In this case the states are the integers, both positive and negative. The state represents the position of the random walk at the given time.



The transition probabilities are:

$$p_{ij} = \begin{cases} p & \text{if } \ j = i+1 \quad \text{(move right)} \\ q & \text{if } \ j = i-1 \quad \text{(move left)} \\ 0 & \text{in all other cases} \end{cases}$$

The matrix of this Markov chain is an infinite matrix part of which looks like this:

$$M = \begin{bmatrix} \ddots & & & & & & & \\ & o & p & o & o & o & o & \\ & q & o & p & o & o & o & \\ & o & q & o & p & o & o & \\ & o & o & q & o & p & o & \\ & o & o & o & q & o & p & \\ & & & & & & & \ddots \end{bmatrix}$$

Unlike the coin-tossing manifestation of the Bernoulli process, the powers M^n of this transition matrix are progressive more complicated. Moreover, the behavior of this Markov chain does depend on the initial distribution X_0. Typically X_0 will take some value i with probability 1, in which case we say that i is the <u>starting point</u> of the random walk. If we start at $i = 0$, the successive distributions $X_0, X_1, X_2, \ldots.$ of this Markov chain are:

$$\begin{array}{ll} X_0 & \vec{u}_0 = [\ \cdots\ 0,\ 0,\ \ 1,\ 0,\ \ 0,\ \cdots\ ] \\ X_1 & \vec{u}_1 = [\ \cdots\ 0,\ q,\ \ 0,\ p,\ \ 0,\ \cdots\ ] \\ X_2 & \vec{u}_2 = [\ \cdots\ q^2,\ 0,\ 2pq,\ 0,\ p^2,\ \cdots\ ] \\ & \quad \cdots \end{array}$$

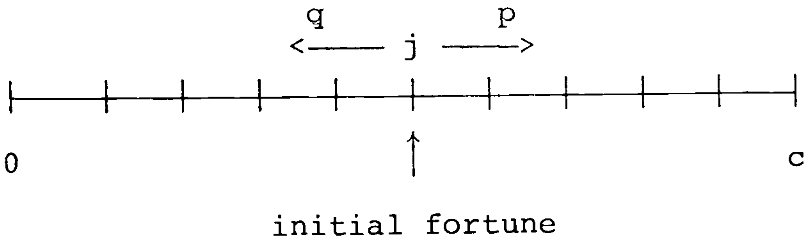
When $p = q = 1/2$ we say the random walk is symmetric. In this case the transition matrix M is symmetric.

As we have already remarked, the random walk model and the coin-tossing model are both interpretations of a single process: the Bernoulli process. However, the two models correspond to very different Markov chains, and hence one asks completely different questions about the two models. For example, we will consider the question of how long it takes for the random walk to return to its starting point. One might also consider how many times the random walk crosses the origin. These questions will be considered not only for random walks but also for more general Markov chains. A great number of physical and chemical phenomena can be modelled using Markov chains and random walks in particular. For example, polymer growth can be modelled using two- and three-dimensional random walks. A two-dimensional random walk is just a pair of independent one-dimensional random walks proceeding simultaneously.

2. The Ruin Problem

Suppose that we are gambling in a casino. Suppose that we bet \$1 on each play and that we win another dollar with probability p and lose the dollar with probability q. This situation is modelled by a random walk. The starting point X_0 is our *initial fortune*, and the state at the time n is our fortune at that time. Unfortunately, the random walk model we have just considered does not take into consideration the fact that we cannot continue playing if we run out of money. Furthermore, there is a number c (possibly very large) such that if we ever succeed in reaching this state the gambling house must stop allowing us to play (or we may simply choose to stop playing if our fortune ever reaches c).

The Markov chain corresponding to this situation is called a *random walk with absorbing barriers*. The *barriers* are the states 0 and c, and these have the property that once one of them occurs, the subsequent states of the Markov chain are all this same state.



This Markov chain has only finitely many states so the transition probability matrix M is an ordinary square matrix. All rows of M except the first and last have the same form as the rows of the barrierless random walk. The top and bottom rows

$$M = \begin{array}{c|ccccccccccc|}
 & 0 & 1 & 2 & 3 & 4 & \cdot & \cdot & \cdot & c-2 & c-1 & c \\
0 & 1 & 0 & 0 & 0 & 0 & & & & & & \\
1 & q & 0 & p & 0 & 0 & & & & & & \\
2 & 0 & q & 0 & p & 0 & & & & & & \\
3 & 0 & 0 & q & 0 & p & & & & & & \\
\cdot & & & & & & \cdot & & & & & \\
\cdot & & & & & & & \cdot & & & & \\
\cdot & & & & & & & & \cdot & & & \\
c-1 & & & & & & & & & q & 0 & p \\
c & & & & & & & & & 0 & 0 & 1
\end{array}$$

have a 1 as the first and last entries respectively, indicating that if either of these states is a starting state, then the ending state is the same state.

Other kinds of barriers are possible. Suppose that if our fortune decreases to zero at any time, we are given a $1 advance (or loan) from an outside source ("Daddy") so that we can continue to play. We call this a reflecting barrier. Still another possibility is the elastic barrier for which we are either reflected or remain in the same state depending on some probability. In other words "Daddy" will give us a loan, but we may have to wait for it. The transition matrix for a random walk having a reflecting barrier at c and an elastic barrier at 0 has this form:

$$\begin{array}{r} \text{elastic barrier} \longrightarrow \\ \\ \\ \\ \\ \\ \end{array} \quad M = \begin{bmatrix} s & r & 0 & & & \\ q & 0 & p & & & \\ & q & 0 & p & & \\ & & & \ddots & & \\ & & & q & 0 & p \\ & & & 0 & 1 & 0 \end{bmatrix} \begin{array}{l} \\ \\ \\ \\ \\ \longleftarrow \text{reflecting barrier} \end{array}$$

A problem of obvious relevance to any gambler is the probability, for a given initial fortune, that the random walk will reach state 0 before reaching state c. If the gambler's fortune ever reaches state zero, we say the gambler is "ruined". For this reason this problem has come to be called the ruin problem. This is only the beginning of the general question of how Markov chains behave in the long run, which we will consider later in this chapter.

Let A be the event "in the random walk with absorbing barriers, the walk reaches 0 before reaching c". Then the ruin problem is to compute $u_j = P(A|X_0 = j)$, for all j. Now u_0 is 1 because $X_0 = 0$ means we are ruined from the start; and u_c is 0 for the opposite reason. For $j \neq 0, c$ we use the conditional law of alternatives (see section V.1) conditioning on the possible values of X_1. There are only two alternatives, $(X_1 = j-1)$ or $(X_1 = j+1)$, when $(X_0 = j)$. Therefore,

$$\begin{aligned} P(A|X_0 = j) &= P(A|(X_0=j) \cap (X_1=j-1))P(X_1=j-1|X_0=j) \\ &+ P(A|(X_0=j) \cap (X_1=j+1))P(X_1=j+1|X_0=j) \\ &= P(A|X_1=j-1)q + P(A|X_1=j+1)p, \end{aligned}$$

by the Markov property and the definition of the transition probabilities. We now make the important observation that in a homogeneous Markov chain we may view any of the random variables X_n as the initial distribution of the sequence $X_n, X_{n+1}, X_{n+2},\ldots.$ which is itself a Markov chain having the same transition matrix as the original Markov chain $X_0, X_1, X_2,\ldots.$. In other words, except for the numbering of the random variables and the initial distribution, this new Markov chain is the same as the old Markov chain. Since A is the event that the gambler is eventually ruined, it does not depend on the numbering of the random variables $X_0, X_1, X_2,\ldots.$. That is, we don't care <u>when</u> the gambler is ruined. Hence

$$P(A|X_1 = j-1) = u_{j-1},$$

and $$P(A|X_1 = j+1) = u_{j+1}.$$

Therefore, $u_j = P(A|X_0 = j) = u_{j-1}q + u_{j+1}p$, for $0 < j < c$.

An equation of the above form is called a <u>difference equation</u>, while the conditions $u_0 = 1$ and $u_c = 0$ are its <u>boundary conditions</u>. A difference equation can be solved in a manner exactly analogous to a differential equation, except that instead of exponential functions $u(x) = e^{\alpha x}$ we use the functions $u_j = \alpha^j$, where α is a constant. We'll proceed by steps to emphasize the similarity with differential equation techniques.

Step 1. Determine the possible values for α.

If we substitute $u_j = \alpha^j$ in the equation $u_j = u_{j-1}q + u_{j+1}p$, we get $\alpha^j = \alpha^{j-1}q + \alpha^{j+1}p$. Dividing by α^{j-1}, we find that $\alpha = q + \alpha^2 p$, a quadratic equation in α. Solving for α we find that

$$\alpha = \frac{1 \pm \sqrt{1-4pq}}{2p} = \frac{1 \pm \sqrt{1-4p+4p^2}}{2p} = \frac{1 \pm (1-2p)}{2p} = \left\{\frac{q}{p}, 1\right\} .$$

Notice that there are two cases. When $p = q = 1/2$, there is a double root $\alpha = 1$; and when $p \neq q$, there are two distinct roots.

Step 2. Find the general solution to the difference equation.

When there are distinct roots, the general solution is just an arbitrary linear combination of the functions α^j as α ranges over all roots. Thus when $p \neq q$, the general solution is

$$\begin{aligned} u_j &= C_1 \left(\frac{q}{p}\right)^j + C_2 (1)^j \\ &= C_1 \left(\frac{q}{p}\right)^j + C_2 . \end{aligned}$$

On the other hand, if there are multiple roots we must use functions of the form α^j, $j\alpha^j$, $j^2\alpha^j$,.... using as many as the multiplicity of α as a root. Therefore the general solution when $p = q = 1/2$ is

$$\begin{aligned} u_j &= C_1(1)^j + C_2 \cdot j(1)^j \\ &= C_1 + C_2 j \ . \end{aligned}$$

Step 3. Use the boundary conditions to find the particular solution.

The boundary conditions are $u_0 = 1$ and $u_c = 0$. So when $p \neq q$ we have:

$$\begin{aligned} u_0 &= 1 = C_1\left(\frac{q}{p}\right)^o + C_2 = C_1 + C_2 \\ u_c &= 0 = C_1\left(\frac{q}{p}\right)^c + C_2 . \end{aligned}$$

Solving for C_1 and C_2 we find that

$$\begin{aligned} C_1 &= 1/(1 - (q/p)^c) \\ C_2 &= -(q/p)^c/(1 - (q/p)^c) . \end{aligned}$$

Hence the particular solution we seek is

$$u_j = \frac{(q/p)^j - (q/p)^c}{1 - (q/p)^c} \ .$$

On the other hand, when $p = q = 1/2$, we have

$$\begin{aligned} u_0 &= 1 = C_1 + C_2 \cdot 0 = C_1 \\ u_c &= 0 = C_1 + C_2 \cdot c \ . \end{aligned}$$

Solving for C_1 and C_2 we find that

$$C_1 = 1$$

$$C_2 = -1/c.$$

Hence the solution in this case is

$$u_j = 1 - j/c.$$

Summarizing, we find that the probability of ruin starting from the initial fortune j is

$$P(A|X_0=j) = \begin{cases} 1 - j/c, & \text{if } p = q = 1/2 \text{ (the game is fair)} \\ \dfrac{(q/p)^j - (q/p)^c}{1 - (q/p)^c}, & \text{if } p \neq q \text{ (the game is unfair)} \end{cases}$$

The Solution to the Ruin Problem

The so-called gambler's ruin paradox refers to the fact that the above probabilities are very close to 1 when perfectly reasonable values of p, q, j and c are used. For example, suppose that a gambler has an initial fortune of \$500. Suppose that the gambler decides to be smart and will quit the moment his fortune reaches \$1000. He is playing \$1 bets on black or red in the game of roulette. In this game $p = 18/38$ and $q = 20/38$. He reasons that although the game is unfair, the odds against his eventual win are only 10:9. This would be true if he bet his entire \$500 on one turn of the wheel. However, by betting only \$1 at a time his probability of ruin is, by the above formula,

$$P(A|X_0=500) = \frac{(20/18)^{500} - (20/18)^{1000}}{1 - (20/18)^{1000}}$$

$$= 1 - \frac{(20/18)^{500} - 1}{(20/18)^{1000} - 1}$$

$$\simeq 1 - (10/9)^{-500}$$

$$> 1 - 10^{-22}.$$

Therefore, the gambler has less than one chance in 10^{22} of eventually winning!

On the other hand, this says nothing about how long it will take for the gambler to be ruined nor whether the gambler will enjoy occasional "winning streaks". One can clearly see that it will take many more than 500 turns of the wheel on the average before the gambler is ruined. Moreover, one can show that "winning streaks" and "losing streaks" (when suitably defined) are actually probable events during long betting sessions. So the "structure" of the gambler's ruin is much more complicated than the solution to the ruin problem suggests. It is this complexity that the gambler is presumably paying for when he bets smaller bets instead of the one grand $500 bet on a single turn of the wheel.

We end by considering what happens when $c \to \infty$. One can think of this as a random walk with just one absorbing barrier. It corresponds to the gambling situation in which the house has infinite resources, and the gambler sets no limit on how much he is willing to win. There are three cases.

<u>Unfair game to the gambler</u>: $p < q$. In this case,

$$u_j = \frac{(q/p)^j - (q/p)^c}{1 - (q/p)^c} = \frac{(p/q)^{c-j} - 1}{(p/q)^c - 1} \longrightarrow \frac{0 - 1}{0 - 1} = 1$$

as $c \to \infty$, because $p/q < 1$. So the gambler certainly loses in this case. This is no surprise.

<u>Fair game</u>: $p = q = 1/2$. In this case,

$$u_j = 1 - j/c \longrightarrow 1 \quad \text{as} \quad c \to \infty.$$

Therefore, the gambler eventually loses even in a fair game.

<u>Unfair game to the house</u>: $p > q$. In this case $q/p < 1$ so

$$u_j = \frac{(q/p)^j - (q/p)^c}{1 - (q/p)^c} \longrightarrow (q/p)^j.$$

Hence there is a positive probability that the gambler continues winning forever. This follows essentially from the fact that $p > q$ produces a "drift" of the random walk to the right as if there were a force acting in the positive direction.

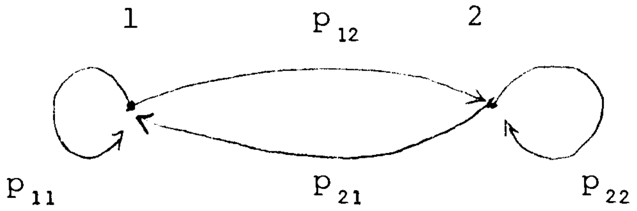
3. The Graph of a Markov Chain

The graph of a homogeneous Markov chain is an effective method of describing and picturing a Markov chain. Moreover, by using these graphs we can view all homogeneous Markov chains as being "random walks" but on the graph rather than on a straight line.

Let us begin with a simple example. This is a simple model of machine operation. We suppose that there are two states 1 = "the machine runs" and 2 = "the machine is broken down". During each unit of time (say every hour), the machine either works or doesn't work. There is a certain probability p_{11} that a working machine will stay working and a probability p_{22} that a broken machine will still be broken. If we assume that these apply to the machine during each unit of time independently of previous states, then this model is a homogeneous Markov chain. Its transition matrix is

$$\begin{bmatrix} p_{11} & p_{12} = 1-p_{11} \\ p_{21} = 1-p_{22} & p_{22} \end{bmatrix}.$$

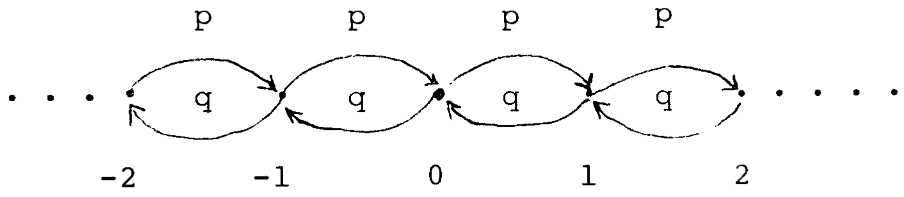
To picture this Markov chain we draw two points (vertices) to represent the states. We then draw lines (edges) between these vertices, with arrowheads to denote direction, indicating the possiblity of passage from one state to another (or the same) state.



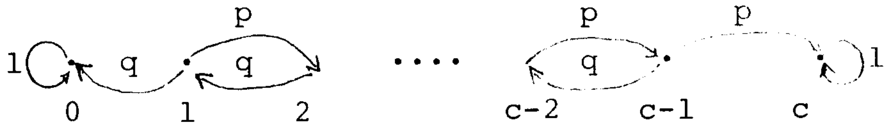
The Graph of a two-state Markov chain.

We now think of the model as describing the motion of a point along the edges of this graph. During each unit of time, the point follows exactly one of the edges in the indicated direction to its other end. The label of the edge denotes the probability that that edge will be chosen. We may think of the point as representing the position of someone "walking" on the graph in which case our model represents a "random walk on the graph."

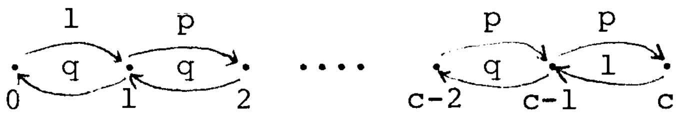
As graphs, the various random walks we considered in the last section look like the following:



Random walk (no barriers)

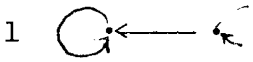


Random walk (absorbing barriers)

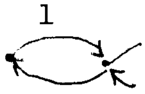


Random walk (reflecting barriers)

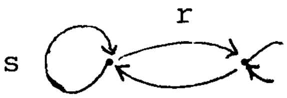
The only features that change in the above models are the boundaries. There are three kinds of boundaries:



absorbing boundary



reflecting boundary



elastic boundary

Definition The graph of a homogeneous Markov chain consists of

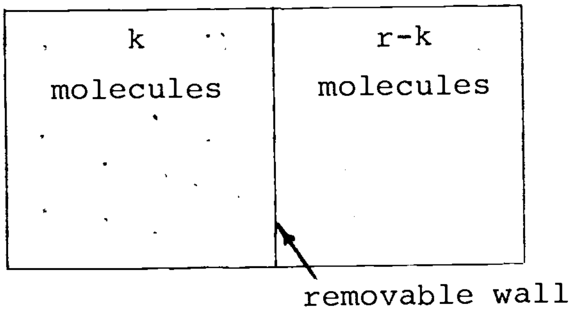
(a) one vertex for every state,

(b) for every pair of states i and j such that $p_{ij} \neq 0$, a directed edge from state i to state j.

Notice that we do not have an edge from state i to state j if $p_{ij} = 0$.

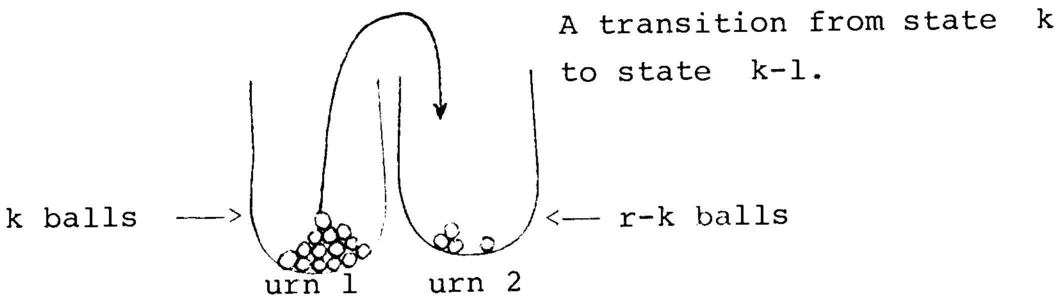
The Ehrenfest Diffusion Model

The Ehrenfest model attempts to explain the following physical experiment. A container is divided into two equal parts by a removable wall. We place k gas molecules in the one part and $r - k$ in the other. Then we remove the wall and wait for a time. If we



now reinsert the wall, we will find almost the same number of particles in each part no matter how many particles were initially placed in the two parts. To find an explanation for this phenomenon is called the diffusion problem.

This model was one of the earliest successful attempts to explain the phenomenon of diffusion using probability. Much more sophisticated models now exist, but it is best to start with the simplest model. For this model we imagine that we have two urns or containers filled with r balls or particles, k in urn 1 and r-k in urn 2. The state of the model is

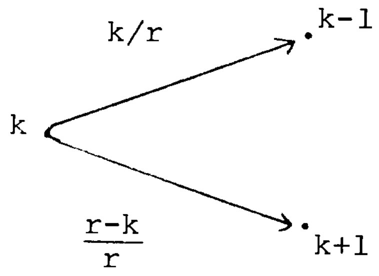


the number of particles in urn 1. A transition of the model consists of transferring one particle from one urn to the other urn. Therefore, there are two possible transitions from state k: to state k-1 or to state k+1. The transition probabilities are assigned in such a way that every particle has the same probability of being transferred to the other urn as any other particle. Therefore the transition probabilities are:

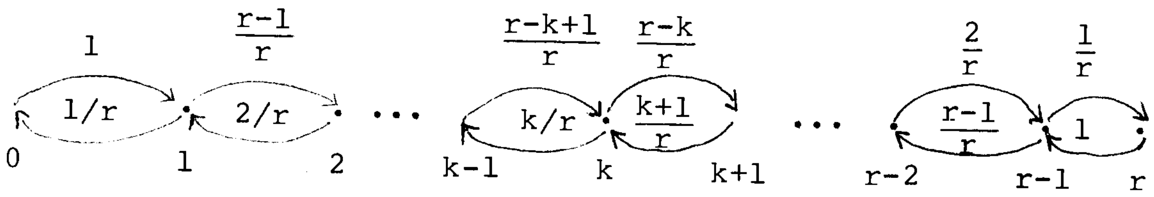
$$p_{k,k-1} = k/r$$

$$p_{k,k+1} = (r-k)/r$$

or using graphs the transitions from k look like this:



The entire graph of this Markov chain is

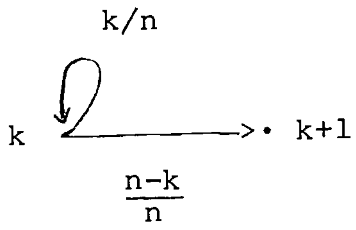


The graph of the Ehrenfest Diffusion Model

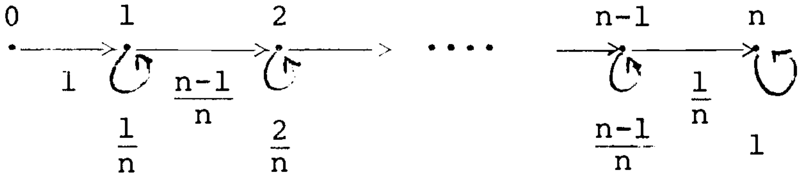
In other words, this Markov chain is a random walk with reflecting barriers but with a "central force" tending to keep the state near $\frac{r}{2}$. We will consider in section 5 what it means to say that the state of this model "tends" to be near $r/2$.

Balls into Boxes

Problems of placing balls into boxes can often be stated in terms of Markov chains. For example, suppose we are sequentially placing balls into n boxes and that we want to know how fast the boxes are being "filled". The state of this Markov chain is the number of boxes having at least one ball. The transitions are either from a state k to the same state if the next ball goes into an already occupied box or to the state $k+1$ if the next ball occupies a new box. Since each box is equally likely to contain the next ball, the probabilities for these two cases are k/n and $1-k/n$ respectively. The graph of this Markov chain looks like this:



The transitions from state k in the Balls into Boxes Markov chain



The graph of the Balls into Boxes Markov Chain

A Genetics Model

The laws of genetics in biology are intrinsically probabilistic. We will consider a very simplified model but one which exhibits the basic ideas. We imagine that a relatively small population of females is introduced to a large ambient population. Consider a single gene having two alleles: a dominant allele A and a recessive allele a. Suppose that the distribution of the three possible genotypes is [p, q, r] in the ambient population, i.e. the fraction of the ambient population having genotype AA is p, having Aa is q and having aa is r. If the females mate the males randomly (at least with respect to this gene), then the distribution of the genotypes in successive generations of females in the subpopulation will form a Markov chain.

The states of the Markov chain are the three genotypes, and the transitions consist of the change of state from a mother to her daughter. The probabilities for parents having given genotypes are the well-known Mendelian laws:

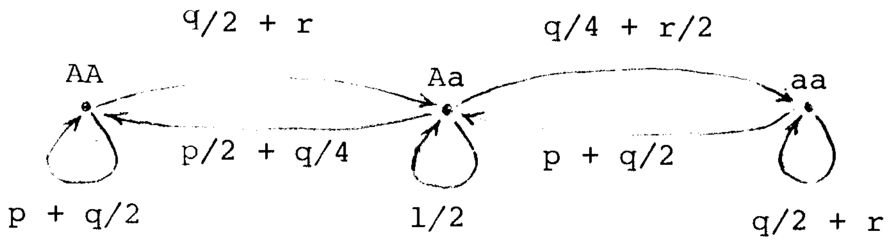
parents		children		
mother	father	AA	Aa	aa
AA	AA	1	0	0
AA	Aa	1/2	1/2	0
AA	aa	0	1	0
Aa	AA	1/2	1/2	0
Aa	Aa	1/4	1/2	1/4
Aa	aa	0	1/2	1/2
aa	AA	0	1	0
aa	Aa	0	1/2	1/2
aa	aa	0	0	1

The Mendelian probabilities for a pair of alleles of one gene.

Since we know the distribution of the genotypes in the ambient population, and since we have assumed the females mate randomly with respect to this gene, we can compute the probabilities for each given female genotype to give rise to a given daughter genotype:

mother	daughter		
	AA	Aa	aa
AA	p + q/2	q/2 + r	0
Aa	p/2 + q/4	1/2	q/4 + r/2
aa	0	p + q/2	q/2 + r

This is the transition matrix of the Markov chain. The graph of the Markov chain is:



We will leave it as an exercise to alter this model to include "preferences" of females of a given genotype for males of another (or the same) genotype as well as to include "survival probabilities" for each of the genotypes of the daughters. One can also construct a model that includes the variation of the distribution of the genotypes of both sexes. The resulting model is a pair of interacting Markov chains acting simultaneously.

Unlike our other examples of Markov chains, we have not considered this Markov chain as a "random walk." We could do so by considering only one "line" of females: a mother, her oldest daughter, her oldest daughter, etc. But as long as the number of children born by a female is independent of her genotype, it is more reasonable to regard this Markov chain as the sequence of distributions of the successive female generations.

More generally, we can view any Markov chain not as a random walk by a single particle but as a random walk by a very large population of particles all simultaneously "walking" on the graph. It is this point of view that is best when we consider the long term behavior of a Markov chain. The behavior of a single particle along its walk can be quite intricate. But the general behavior of the whole population of points is very predictable and stable.

4. The Markov Sample Space

We have seen three definitions of a Markov chain so far. We first defined it to be a sequence $X_0, X_1, X_2, \ldots.$ of random variables satisfying the Markov property. We then saw that it is equivalent to specify a stochastic matrix and a stochastic vector. Finally we saw that we can visualize Markov chains as random walks on graphs. But a Markov chain is a stochastic process so we must have a sample space and a probability.

The sample space Ω of a Markov chain consists of all possible infinite paths along edges of its graph. By "infinite" we mean that the path has a starting point but no ending point. In other words, the sample space is the set of all possible sequences $(i_0, i_1, i_2, \ldots.)$ of states. This formulation is formally analogous to the definition of the Bernoulli process (coin tossing) which is a special case.

To define the probability on Ω we need the transition matrix M and the stochastic vector $\vec{u}_0$ representing the initial distribution. It is standard notation to write $p_{ij}^{(n)}$ for the entries of the matrix M^n; $p_{ij}^{(n)}$ is the probability for the Markov chain to be in state j given that it was in state i exactly n units of time previously. The components of $\vec{u}_0$ are usually denoted by a_i. The elementary events of the sample space Ω are the subsets $(X_n = i)$, i.e. the set of all paths whose n^{th} vertex corresponds to state i. The probability of an elementary event is given by

$$P(X_n = i) = \sum_j a_j p_{ji}^{(n)} .$$

So far this process is defined analogously to the Bernoulli process. However, we do not postulate that elementary events are independent. Instead we specify their probabilities to be

$$P((X_{n_1} = i_1) \cap (X_{n_2} = i_2) \cap \cdots \cap (X_{n_k} = i_k))$$

$$= \sum_j a_j p_{ji_1}^{(n_1)} \; p_{i_1 i_2}^{(n_2-n_1)} \cdots \; p_{i_{k-1} i_k}^{(n_k-n_{k-1})} .$$

In particular, we define

$$P((X_0 = i_0) \cap (X_1 = i_1) \cap \cdots \cap (X_n = i_n))$$

$$= a_{i_0} p_{i_0 i_1} \; p_{i_1 i_2} \cdots \; p_{i_{n-1} i_n} .$$

This last expression determines the preceding ones.

Definition Given a stochastic matrix $M = (p_{ij})$ and a stochastic vector $\vec{u}_0 = (a_i)$, the Markov chain whose transition probability matrix is M and whose initial distribution is $\vec{u}_0$ is defined by

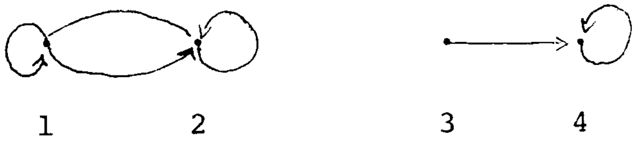
(1) the sample space Ω is the set of all possible sequences of states $(i_0, i_1, \ldots.)$;

(2) the elementary events are the subsets $(X_n = i)$ of all sequences whose n^{th} entry is i;

(3) the probability is defined by
$P((X_0=i_0)\cap(X_1=i_1)\cap\cdots\cap(X_n=i_n))=a_{i_0}p_{i_0i_1}p_{i_1i_2}\cdots p_{i_{n-1}i_n}$.

Connectivity

We now classify Markov chains with respect to various properties relevant to their long term behavior. The most obvious property is connectivity. If we draw the graph of the Markov chain, it should be clear what



A Markov chain having two connected parts

we mean when we say the graph is disconnected: it is made up of two or more parts having no edges between the parts. If the graph is not disconnected, we say the Markov chain is connected. Clearly each connected part of a disconnected Markov chain acts like a Markov chain by itself, independent of the rest of the Markov chain. Because of this we will always assume our Markov chains are connected.

Persistence and Transience

The next property we consider is recurrence. Having once occurred, a state must occur again or it may not. More precisely, let A_i be the event "state i eventually occurs at some time after 0." Then either $P(A_i | X_0 = i) = 1$ or $P(A_i | X_0 = i) < 1$. We call those two possibilities persistence and transience.

Definition A state i is persistent if the probability of returning to state i (after it has occurred at least once) is 1. A state i is transient if the probability is positive for the state i never to occur again.

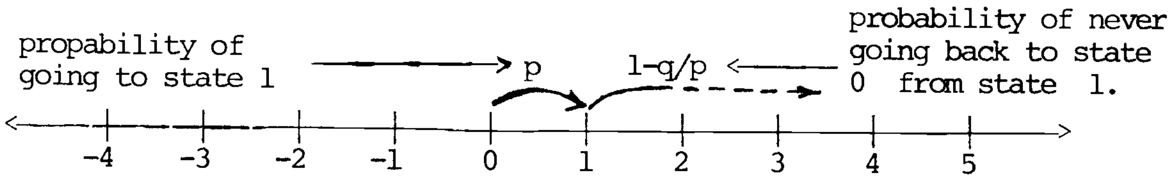
Now, having once occurred, a persistent state must necessarily occur infinitely many times: each time it occurs, we repeat the argument that it must occur once more, so it can never "stop" occurring.

Consider the various random walks we have seen so far. In the random walk with absorbing barriers, the barriers are obviously persistent. Notice that in this case only one of the persistent states can occur during any one "walk". Although persistent states occur infinitely often if they ever occur at all, it is quite possible in a connected Markov chain for a persistent state never to occur. The interior states in the random walk with absorbing barriers are all transient because we know with probability 1 that either one or the other barrier will be encountered eventually. In the random walk with one absorbing barrier and one reflecting (or elastic) barrier, there is just one persistent state. On the other hand, if both barriers are reflecting or elastic, all the states become persistent. We leave it as an exercise to prove these last two statements using the solution to the ruin problem.

Finiteness

With respect to persistence and transience, there is a striking difference between finite and infinite Markov chains. In a finite Markov chain some state must be persistent. But for infinite Markov chains it is quite possible for every state to be transient. Consider the ordinary barrierless random walk. Suppose that the random walk is not symmetric, say $p > q$. Recall that in our solution to the ruin problem, we noted that there is a positive probability, $1 - (q/p)^j$, such that, starting in state j, the random walk forever drifts to the right and never encounters state 0.

Now, if we start in state 0, the next state is state 1 with probability p, and from here the probability is 1 - q/p for 0 never to occur again. Therefore with probability $p(1-q/p) = p - q > 0$, state 0 never occurs again. By the same argument all the states of a nonsymmetric barrierless random walk are transient. In the long run, a nonsymmetric random walk drifts forever either to the right if $p > q$ or to the left if $p < q$.

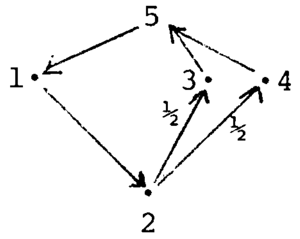

propability of going to state 1
p
1-q/p
probability of never going back to state 0 from state 1.
-4
-3
-2
-1
0
1
2
3
4
5

The probability of never returning to state 0 is at least p-q.

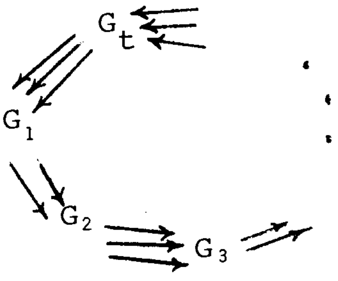
On the other hand, for the symmetric random walk every state is persistent. In our solution to the ruin problem, we noted that starting in state $j > 0$, state 0 eventually occurs with probability 1, and this generalizes to any two states i and j. So not only is every state persistent, but also every state occurs infinitely often.

Periodicity

The last property we will consider is periodicity. An example of a periodic Markov chain is the following:



We will give a precise definition later. Such a Markov chain will "cycle" endlessly with period 4 in a merry-go-round fashion. These are not very interesting Markov chains with respect to long term behavior since they all essentially look more or less like this one. More precisely, one can prove that one can divide the states into classes $G_1, G_2, \ldots, G_t$ in such a way that the graph of the Markov chain looks like this:



The only edges in the graph are from states in G_i to states in G_{i+1} or from states in G_t to states in G_1.

And moreover, the long term behavior of each piece G_i is just as if it were a Markov chain by itself. For this reason we will assume that our Markov chains are not periodic.

Ergodicity

Having gone through all the above preliminaries we find that the most interesting Markov chains with respect to their long term behavior are finite, connected and nonperiodic. We will assume in addition that every state is persistent.

The reason for this is that in a finite Markov chain no transient state can occur more than finitely many times. Hence all transient states eventually cease to be relevant.

Definition A homogeneous Markov chain is said to be ergodic if it is finite, connected and nonperiodic, and if all its states are persistent.

For the rest of this chapter we will study only ergodic Markov chains.

5. Steady States of Ergodic Markov Chains

The most surprising fact about ergodic Markov chains is that the long term behavior of such a Markov chain is independent of the initial distribution X_0. That is, no matter what the initial distribution X_0, the distributions of the X_n's as $n \to \infty$ will tend toward one particular distribution which we call the steady state or invariant distribution.

The best way to view this distribution is take the point of view mentioned at the end of section 3. Instead of thinking of the Markov chain as the motion of a single particle along the graph, we think of an entire population of particles as simultaneously "walking" along the graph. The steady state distribution has the property that although individual particles are in constant motion, the population as a whole has a fixed distribution. So if we choose not to distinguish one particle from another, we would perceive no change as the Markov chain proceeds in time.

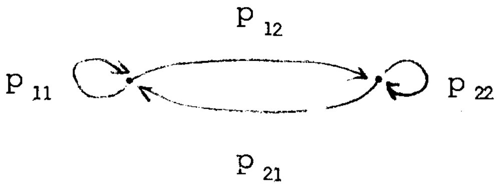
Definition For any Markov chain $X_0, X_1, X_2, \ldots,$ a probability distribution for X_0 such that all the X_i's are equidistributed is called a steady state or invariant distribution of the Markov chain.

To find a steady state distribution we make use of the terminology of vectors and matrices that we introduced in section 1. If we write $\vec{u}_0$ for the vector corresponding to the initial distribution and if M is the transition matrix, then

$$\vec{u}_1 = \vec{u}_0 M$$

is the vector corresponding to the distribution of X_1. Now if $\vec{u}_1 = \vec{u}_0$, then clearly all subsequent distributions will be the same as the first two. Hence a steady state distribution corresponds to an eigenvector whose eigenvalue is 1. Finding all such eigenvectors for a given matrix M is a simple exercise in linear algebra (simultaneous equations). Having found all such eigenvectors, the steady state distributions are those whose components are between 0 and 1 and add to 1.

Consider for example the machine operation model



whose transition matrix is $\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$, where $p_{12} = 1 - p_{11}$ and $p_{21} = 1 - p_{22}$. To find a steady state distribution we must solve

$$[x_1, x_2] = [x_1, x_2] \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

or $\begin{Bmatrix} x_1 = x_1 p_{11} + x_2 p_{21} \\ x_2 = x_1 p_{12} + x_2 p_{22} \end{Bmatrix}$ for x_1 and x_2. This system of equations reduces to the single equation

$$x_2 = x_1 \cdot \frac{p_{12}}{p_{21}}$$

Therefore the general eigenvector belonging to the eigenvalue 1 is

$$C[1, p_{12}/p_{21}].$$

This will be a stochastic vector provided

$$C + Cp_{12}/p_{21} = 1.$$

Solving for C, we find that the invariant distribution of this Markov chain is $\left[\frac{p_{21}}{p_{12} + p_{21}}, \frac{p_{12}}{p_{12} + p_{21}}\right]$

or in terms of X_0,

$$P(X_0 = 1) = \frac{p_{21}}{p_{12} + p_{21}}$$

$$P(X_0 = 2) = \frac{p_{12}}{p_{12} + p_{21}} .$$

For example, if the machine breaks down with probability 1/10 every hour and if a broken machine will go back into service with probability 1/2 every hour, then the machine will be running

$$\frac{p_{21}}{p_{12} + p_{21}} = \frac{.5}{.1 + .5} = \frac{.5}{.6} = \frac{5}{6}$$

of the time. Moreover, this will be true in the long run whether the machine is initially running or initially broken down.

Waiting Times and the Recurrence Theorem

In all the stochastic processes we have studied so far, the waiting times have played a crucial role. So it is also with Markov chains. For each state j, we define the random variable T_j to be the waiting time for return to state j, given that one starts in state j. More generally we have the waiting times T_{ij} for the occurance of state j starting from state i. In terms of Markov events

$$(T_{ij}=n) = (X_0=i) \cap (X_1 \neq j) \cap \cdots \cap (X_{n-1} \neq j) \cap (X_n=j).$$

Of course, we have that $T_j = T_{jj}$. For Markov chains in general these are not really random variables because one may have states for which $\sum_n P(T_{ij}=n) \neq 1$ (in fact one can have $P(T_{ij}=n) = 0$ for all n). We call such an object a defective random variable. However, we specifically chose to restrict attention to ergodic Markov chains because in this case all the waiting times are ordinary random variables.

There is a slight technicality that we ought to mention briefly. The waiting times T_{ij} are not defined on the same sample space. In fact T_{ij} is defined on the Markov chain for which X_0 takes initial value i with probability 1.

The standard notation for $P(T_j=n)$ is $f_j^{(n)}$ and for $P(T_{ij}=n)$ is $f_{ij}^{(n)}$. The random variable T_j is also called the recurrence time of state j, and its expectation $E(T_j) = \sum_n nf_j^{(n)}$ is called the mean recurrence time of state j. We can now give a precise definition of what it means for a Markov chain to be ergodic.

Definition A finite, connected Markov chain is ergodic if

(1) for every state j, $\sum_n P(T_j=n) = 1$ (every state is persistent),

(2) for every state j, $p_{jj}^{(n)} > 0$ except for at most a finite number of times n (no state is periodic).

The most important facts about waiting times are contained in the following remarkable

Recurrence Theorem For any finite, ergodic Markov chain,

(1) $p_{ij}^{(n)} \longrightarrow 1/E(T_j)$ as $n \longrightarrow \infty$ for all i and j,

(2) the components of the steady state distribution are $1/E(T_j)$.

One can rewrite statement (1) in the form

$$\lim_{n\to\infty} P(X_n = j) = 1/E(T_j).$$

Combining this with statement (2), we find that ergodic Markov chains satisfy an analogue of the Central Limit Theorem:

No matter what the initial distribution is, the distributions of the random variables X_n of an ergodic Markov chain necessarily converge to the steady state distribution.

Furthermore, the steady state distribution may be regarded as specifying the average time the random walk exists in the various states. We then have that the average length of time between occurrences of state i is the inverse of the average time spent in state i. While this is an intuitively clear result, it is far from being easy to prove.

The Ehrenfest Diffusion Model

We illustrate the Recurrence Theorem for a nontrivial example. In this model the transition probabilities are

$$p_{ij} = \begin{cases} i/r & \text{if } j = i - 1 \\ 1 - i/r & \text{if } j = i + 1 \\ 0 & \text{otherwise} \end{cases}$$

We begin by computing the steady state distribution. We must solve the system of equations

$$p_j = \sum_i p_i p_{ij} = \begin{cases} p_1/r, & \text{if } j = 0 \\ p_{r-1}/r, & \text{if } j = r \\ p_{j-1}\left(\frac{r-j+1}{r}\right) + p_{j+1}\left(\frac{j+1}{r}\right), & \text{otherwise} \end{cases}$$

The third equation gives us a recursive expression for p_j

in terms of p_{j-1} and p_{j-2}:

$$p_j = p_{j-1}\left(\frac{r-j+1}{r}\right) + p_{j+1}\left(\frac{j+1}{r}\right),$$

$$p_{j+1}\left(\frac{j+1}{r}\right) = p_j - p_{j-1}\left(\frac{r-j+1}{r}\right),$$

$$p_{j+1} = p_j\left(\frac{r}{j+1}\right) - p_{j-1}\left(\frac{r-j+1}{j+1}\right),$$

shifting indices:

$$p_j = p_{j-1}\left(\frac{r}{j}\right) - p_{j-2}\left(\frac{r-j+2}{j}\right).$$

Now using the fact that $p_0 = p_1/r$, we solve by setting $p_0 = 1$ (the solution is only determined up to a scalar multiple so it doesn't matter what we use for p_0) and by applying the above recursion successively. This gives:

$$p_0 = 1$$

$$p_1 = r$$

$$p_2 = r \cdot \frac{r}{2} - 1 \cdot \frac{r}{2} = \frac{r(r-1)}{2}$$

$$p_3 = \frac{r(r-1)}{2} \cdot \frac{r}{3} - r \cdot \frac{(r-1)}{3} = \frac{r(r-1)(r-2)}{2 \cdot 3}$$

A pattern is clearly developing. It seems that $p_j = \binom{r}{j}$. In fact, the formula

$$p_j = p_{j-1}\left(\frac{r-j+1}{r}\right) + p_{j+1}\left(\frac{j+1}{r}\right)$$

is just a combination of indentities (1) and (5) in section II.5.

Therefore the eigenvectors belonging to eigenvalue 1 of the transition matrix for this Markov chain has j^{th} component $C\binom{r}{j}$, for any constant $C \neq 0$. For the steady state distribution we must choose C so that $\sum_j C\binom{r}{j} = 1$. But we know that $\sum_j \binom{r}{j} = 2^r$ (identity (10) of section II.5). Therefore $C = 2^{-r}$. The steady state distribution therefore has

$$P(X_0 = j) = \binom{r}{j}2^{-r}.$$

This is none other than the binomial distribution for r tosses of a fair coin! In other words, it is as if we placed the particles of the Ehrenfest model into the two urns one at a time according to the toss of a fair coin.

We know that the binomial distribution is closely approximated by the normal distribution. Therefore X_0 is very close to having the distribution $N(r/2\ ,\ r/4)$ $(p = q = 1/2$ and $n = r)$. Therefore $\dfrac{X_0 - r/2}{\sqrt{r/4}}$ is approximately $N(0,1)$. Suppose that we use confidence level 99.9%. Then

$$P\left(-3.3 \leq \frac{X_0 - r/2}{\sqrt{r/4}} \leq 3.3\right) = .999$$

implies that $P\left(|X_0 - r/2| \leq 1.65\sqrt{r}\right) = .999$. In an actual experiment, r will be of the order 10^{24}. Consider the case $r = 10^{24}$. If we replace the barrier between the urns, we will find with probability .999 that

$$|X_0 - 5 \times 10^{23}| \leq 1.65 \times 10^{12} .$$

Although 1.65×10^{12} is a very large number, it is less than 10^{-11} of the number of particles in either urn, so that any departure of X_0 from being exactly $r/2$ would be very difficult to detect.

The value of $P(X_0 = j)$ when X_0 has the steady state distribution is $1/E(T_j)$ by the Recurrence Theorem. Again using the normal approximation, we have that

$$E(T_j) \simeq \frac{\sqrt{2\pi r}}{2} e^{2(j-r/2)^2/r} .$$

If $j = r/2$, we get $E(T_j) \simeq \frac{\sqrt{2\pi r}}{2}$. When $r = 10^{24}$, this is about 10^{12}. This may seem to be very large, but this is only because the unit of time we are using is very small. On the other hand, if $j = 0$ and $r = 10^{24}$, then $E(T_0) = 1/P(X_0 = 0) = 2^r = 2^{10^{24}} \simeq 10^{10^{23}}$. Even if our time scale is as small as 10^{-50}sec, this waiting time dwarfs even the waiting time mentioned in section IV.7 (on the writing of Hamlet at random by a monkey). Clearly this state does not occur very often.

6. Exercises for

Chapter VIII Markov Chains

1. Compute the probability of the gambler's ruin for a gambler having initial fortune \$500 and upper limit on winnings \$1000, who is playing roulette and who is making bets of \$10 on red or black; do the same for bets of \$100. What advice would you give to the gambler? to the gambling casino?

2. In doing homework problems each success improves the chance of another success, while each failure tends to increase the chance of subsequent failure. Build a Markov chain model for this.

3. Consider the following model of the spread of disease. There are N persons in the population. Some are sick and the rest are not.

   (a) when a sick person meets a healthy one, the healthy one becomes sick with probability α,

   (b) all encounters are between pairs of persons,

   (c) all possible encounters in pairs are equally likely,

   (d) one such encounter occurs per unit time,

   (e) during each unit of time each sick person recovers with probability β independently of (a) - (d) and of the pervious time spent sick.

Let X_n be the number of sick persons at time n. Write the transition matrix for this Markov chain, and draw its graph.

4*. Alter the genetics model in section 3 to include a genetic advantage for one of the genotypes (say, for the Aa genotype). How would you include a perference by females having certain genotypes for males having certain other genotypes. Assume that the genotypes AA and Aa are indistinguishable from one another.

5*. Prove that in a finite random walk without absorbing barriers, all states are persistent.

6. A man has two girl friends A and B, one living uptown and one living downtown, respectively. He either visits one of his girl friends on a given evening, or he stays at home. The day after an evening at home he goes to the bus stop at a random time and takes whichever bus comes first, the bus uptown or the bus downtown, visiting A or B, respectively. The buses run in both directions during every 15 minute interval, on a fixed schedule. The man is not too compatiblé with A, for after a visit to her, he stays home the next evening with probability 9/10 and visits her again with probability 1/10. On the other hand, he is quite compatible with B; after a visit to her, he visits her again the next evening with probability 9/10 and stays home with probability 1/10. Set up the Markov chain for this process. Much to the man's surprise, he spends as many evenings with A as with B on the average. Compute how frequently he spends his evening at home, on the average. See exercise 11.16.

7. Compute the steady state distribution of the genetics model in section 3. Notice that it is not in general the same as the distribution of genotypes in the larger population.

8*. Compute the steady state distribution for the more general genetics models in exercise 4.

9. Compute the steady state of the symmetric finite Markov chain with reflecting barriers. How often does "Daddy" advance you a loan on the average?

10. In exercise III.2 the San Francisco bar in question is 100 yards uphill from the Bay, but the drunk's home is only 10 yards uphill from the bar. How probable is it that the drunk falls into San Francisco Bay before finding his way home?

11. In a chemical solution there are initially N molecules, each being one of types A, B, C or D. During every unit of time exactly one collision occurs between a pair of these molecules, all possible collisions being equally likely. During such a collision nothing happens unless the colliding molecules are A and B or are C and D. If A and B collide, there is a probability α that they react and become a pair of C and D molecules. If C and D collide, there is a probability β that they become A and B. In chemical symbols:

$$A + B \underset{\beta}{\overset{\alpha}{\rightleftharpoons}} C + D.$$

The state of this system is totally determined by the number of A molecules. Let the number of A molecules at time n be X_n. What is the transition matrix for this Markov chain? What is the steady state number of molecules of each kind?

12. Same as exercise 11 above, but for the autocatalytic reaction

$$A + A \underset{\beta}{\overset{\alpha}{\rightleftharpoons}} B + A.$$

13*. Generalize exercises 11 and 12 to an arbitrary number of initial molecules of each kind. Can the reaction rate constants be determined from the steady state distribution?

INDEX