Dice, used with a plural verb, means small cubes marked with one to six dots, used in gambling games. Dice, used with a singular verb, means a gambling game in which these cubes are used. Dice (plural) can also refer to any small cubes, especially cube-shaped pieces of food (Cut the cheese into dice).

– MSN Encarta

Lecture VIII
QUICK PROBABILITY

We review the basic concepts of probability theory, using the axiomatic approach first expounded by A. Kolmogorov. His classic [6] is still an excellent introduction. The axiomatic approach is usually contrasted to the empirical or "Bayesian" approach that seeks to predict real world phenomenon with probabilistic models. Other source books for the axiomatic approach include Feller [3] or the approachable treatment of Chung [2]. Students familiar with probability may simply use this lecture as a reference.

Probability in algorithmics arises in two main ways. In one situation, we have a deterministic algorithm whose input space has some probability distribution. We seek to analyze, say, the expected running time of the algorithm. The other situation is when we have an algorithm that makes random choices, and we analyze its behaviour on any input. The first situation is considered less important in algorithmics because we typically do not know the probability distribution on an input space (even if such a distribution exists). By the same token, the second situation derives its usefulness from avoiding any probabilistic assumptions about the input space. Algorithms that make random decisions are said to be **randomized** and comes in two varieties. In one form, the algorithm may make a small error but its running time is worst-case bounded; in another, the algorithm has no error but only its expected running time is bounded. These are known as **Monte Carlo** and **Las Vegas** algorithms, respectively.

There is an understandable psychological barrier to the acceptance of unbounded worst-case running time or errors in randomized algorithms. However, it must be realized that the errors in randomized algorithms are controllable by the user – we can make them as small as we like at the expense of more computing time. Should we accept an algorithm with error probability of 2^{-99}? In daily life, we accept and act on information with a much greater uncertainty or likelihood of error than this.

More importantly, randomization is, in many situations, the only effective computational tool available to attack intransigent problems. Until recently, the standard example of a problem not known to be in the class P (of deterministic polynomial time solvabale problems), but which admits a randomized polynomial-time algorithm is the **Primality Problem**. Since August 2002, Manindra Agrawal, Neeraj Kayal and Nitin Saxeena, in a major breakthrough, has shown that this problem is in P. The current best algorithm for Primality Testing is $O(n^{7.5})$, so it is still not very practical. Thus randomized primality remains the useful in practice. Note that the related problem of factorization of integers does not even have a randomized polynomial time algorithm.

§1. Axiomatic Probability

All probabilistic phenomena occur in a probabilistic space, which we now formalize (axiomatize).

¶1. Sample space. Let Ω be any non-empty set, possibly infinite. We call Ω the **sample space** and elements in Ω are called **sample points**.

We use the following running examples of sample spaces:

(E1) $\Omega = \{H, T\}$ (coin toss). This represents a probabilistic space where there are two outcomes. Typically we identify these outcomes with the results of tossing a coin – head (H) or tail (T).

(E2) $\Omega = \{1, \ldots, 6\}$ (dice roll). This is a slight variation of (E1) representing the outcomes of the roll of dice, with six possible outcomes.

(E3) $\Omega = \mathbb{N}$ (the natural numbers). This is a significant extension of (E1) since there is a countably infinite number of outcomes.

(E4) $\Omega = \mathbb{R}$ (the real numbers). This is a profound extension because we have gone from discrete space to continuous space.

¶2. Event space. Sample spaces becomes more interesting when we give it some structure to form event spaces.

Let $\Sigma \subseteq 2^\Omega$ be a subset of the power set of Ω. The pair (Ω, Σ) is called an **event space** provided three axioms hold:

(A0) $\Omega \in \Sigma$.

(A1) $A \in \Sigma$ implies that its complement is in Σ, $\Omega - A \in \Sigma$.

(A2) If $A_1, A_2, \ldots$ is a countable sequence of sets in Σ then $\cup_{i \geq 1} A_i$ is in Σ.

We call $A \in \Sigma$ an **event** and singleton sets in Σ are called **elementary events**. Thus the axioms (A0) and (A1) imply that $\emptyset$ and Ω are events (the "impossible event" and "inevitable event"), and the complement of an event is an event. Thus, events are closed under countable unions (by (A2)) and also countable intersections using de Morgan's law:

$$\bigcap_i \overline{A_i} = \overline{\bigcup_i A_i}.$$

Great, a non-event is an event!

In some contexts, an event space is also[1] called a **Borel field** or **sigma field**

We now show two basic ways to construct event spaces:

(i) Construction from generator set: let Ω be any set and $G \subseteq 2^\Omega$. Then there is a smallest (by the Axiom of Choice)

[1] Sometimes, "ring" is used instead of "field".

event space $\overline{G}$ containing G; we call this the event space **generated by** G. For example, let $\Omega = \{1, \ldots, 6\}$ (dice example, E2). If $G = \{\{1,2,3\}, \{3,4,5,6\}\}$ then

$$\overline{G} = \{\emptyset, \{3\}, \{1,2\}, \{1,2,3\}, \{4,5,6\}, \{3,4,5,6\}, \{1,2,4,5,6\}, \{1,2,3,4,5,6\}\}.$$

(ii) Subspace construction: if (Ω, Σ) is an event space and $A \in \Sigma$, then we obtain a new event space $(A, \Sigma \cap 2^A)$, which is the **subspace** of (Ω, Σ) **induced** by A.

Let A and B be events. There are two standard notations for events which we will use. First, we will write "A^c" or "$\overline{A}$" for the **complementary event** $\Omega \setminus A$. Also, we write "AB" for the event $A \cap B$ (why is this an event?). This is called the **joint event** of A and B. Two events A, B are **mutually exclusive** if $AB = \emptyset$.

¶3. Event Spaces in the Running Examples. One choice of Σ is

$$\Sigma = 2^\Omega. \tag{1}$$

We call this the **discrete sample space**, which is the typical choice for countable sample spaces such as the running examples (E1), (E2) and (E3). In example (E2) the event $\{3,4,5,6\} \in \Sigma$ may be read: "the event that roll is at least 3". What is wrong in assuming (1) in all situations? This choice certainly gives an event space. The problem arises when need to assign probabilities to events (see next). When Ω is infinite, the choice (1) admits many events for which it is unclear how to assign probabilities. This issue is severe when Ω is uncountable.

We illustrate the standard way to create an event space for $\Omega = \mathbb{R}$, via a generating set $G \subseteq 2^\Omega$. Let G comprise the half-lines

$$H_r := \{x \in \mathbb{R} : x \leq r\}$$

for each $r \in \mathbb{R}$. This generates an event space $\overline{G}$ that is extremely important. It is called the **Euclidean Borel field** and denoted B^1 or $B^1(\mathbb{R})$. An element of B^1 is called an **Euclidean Borel set**. These sets are not easy to describe explicitly, but let us see that some natural sets belongs to B^1. We first show that singletons $\{r\}$, $r \in \mathbb{R}$, belong to B^1:

Definition of $B^1 = B^1(\mathbb{R})$

$$\{r\} = H_r \cap \bigcap_{n \geq 1} H^c_{r+(1/n)}.$$

Then (A2) implies that any countable set belongs to B^1. Furthermore, any open or closed interval belongs to B^1.

¶4. Probability space. So far, we have described concepts that probability theory shares in common with measure theory. Probability properly begins with the next definition: a **probability space** is a triple

$$(\Omega, \Sigma, \Pr)$$

where (Ω, Σ) is an event space and $\Pr : \Sigma \to [0,1]$ (the unit interval) is a function satisfying the following two axioms:

Measure theory is the foundation integral calculus

(P0) $\Pr(\Omega) = 1$.

(P1) if $A_1, A_2, \ldots$ are a countable sequence of pairwise disjoint events then $\Pr(\cup_{i \geq 1} A_i) = \sum_{i \geq 1} \Pr(A_i)$.

We simply call Σ the probability space when Pr is understood. The number $\Pr(A)$ is the **probability** of A. A **null event** is one with zero probability. Clearly, the empty set $\emptyset$ is a null event. It is important to realize that when Ω is infinite, we typically have null events different from $\emptyset$. We deduce that $\Pr(\Omega - A) = 1 - \Pr(A)$ and if $A \subseteq B$ are events then $\Pr(A) \leq \Pr(B)$.

> The student should learn to set up the probabilistic space underlying any probabilistic analysis. Whenever there is discussion of probability, you should ask: what is Ω, what is Σ? This is especially important since probabilists almost never do this explicitly! For finite sample spaces in which each sample point is an event, Pr is completely specified when we assign probabilities to these (elementary) events.

¶5. Probability in the Running Examples. Recall that we have specified Σ for each of our running examples (E1)–(E4). We now assign probabilities to events.

In example (E1), we choose $\Pr(H) = p$ for some $0 \leq p \leq 1$. Hence $\Pr(T) = 1 - p$. If $p = 1/2$, we say the coin is **fair**. In example (E2), let the probability of an elementary event be $1/6$.

In general, when Ω is a finite set, and $\Pr(A) = |A|/|\Omega|$ for all $A \in \Sigma$, we see that the probabilistic framework is simply a convenient language for counting the number of elements in the sets $A \in \Sigma$. The space $(\Omega, 2^\Omega, \Pr)$ where $\Pr(\omega) = 1/|\Omega|$ for all $\omega \in \Omega$ may called the **uniform probability model** for Ω.

For (E3), we may choose $\Pr(i) = p_i \geq 0$ $(i \in \Omega = \mathbb{N})$ subject to

$$\sum_{i=0}^{\infty} p_i = 1.$$

An explicit example is illustrated by $p_i = 2^{-(i+1)}$, since $\sum_{i=0}^{\infty} p_i = 2\sum_{i=0}^{\infty} 2^{-i} = 1$.

For (E4), it is more intricate to define a probability space. We begin with the Euclidean Borel field $B^1 = (\Omega, \Sigma)$, but use the subspace construction (above) to restrict B^1 to a finite interval $[a, b] \subseteq \mathbb{R}$. The resulting sample space is denoted

$$B^1[a, b] = ([a, b], \Sigma \cap 2^{[a,b]})$$

and it is generated by the intervals $[r, b] = H_r \cap [r, b]$, for all $r \leq c \leq b$. The **uniform probability function** for $B^1[a, b]$ is given by

$$\Pr([r, b]) := (b - r)/(b - a) \tag{2}$$

for all generators $[r, b]$ of $B^1[a, b]$. It is not hard to see that $\Pr(A) = 0$ for every countable $A \in \Sigma$. Thus all countable sets are null events.

We could modify this construction to give a probability function for the original $B^1(\mathbb{R})$ since $B^1(\mathbb{R})$ can be identified with $B^1[-1, 1]$ in a natural way.

¶6. Constructing Probability Spaces. We give a product construction for probability spaces: let $\Sigma_i \subseteq 2^{\Omega_i}$ $(i = 1, 2)$ be sample spaces. Let $\Omega = \Omega_1 \times \Omega_2$ and

$$G = \{A_1 \times A_2 : A_i \in \Sigma_i, i = 1, 2\}.$$

Note that in $A_1 \times A_2$ is empty if A_1 or A_2 is empty. The event space $(\Omega, \overline{G})$ generated by G will be denoted $\Sigma_1 \times \Sigma_2$. If $\Pr_i$ is a probability function for Σ_i, then we get a probability function for $\Sigma_1 \times \Sigma_2$ where

$$\Pr(A_1 \times A_2) = \Pr_1(A_1) \Pr_2(A).$$

We leave it as an exercise to show this Pr can be extended into a probability function for $\Sigma_1 \times \Sigma_2$.

We may denote $\Sigma \times \Sigma$ by Σ^2. For $n \geq 3$, let Σ^n denote the event space $(\Sigma^{n-1}) \times \Sigma$. Using this construction, the simple case $\Omega = \{H, T\}$ leads to the event space for a sequence of n coin tosses. We can also let $n = \infty$ and consider Σ^∞. This corresponds to sequences of unbounded length: for instance, imagine tossing a coin until we see a head.

An important type of sample space is based on "decision trees". Assuming a finite tree, the sample points are identified with identified with leaves of the tree and the sample space is 2^Ω. with the set of leaves below each node. How do we assign probabilities? Let assume that at a node of degree d, the probability of taking any of its child is $1/d$. Then the probability of any path is just the product of the probility of taking each edge of the path.

¶7. Quicksort Example. We can generalize the sample space of decision trees above. Let us consider the probability space of Quicksort. Fix any input to Quicksort with n distinct numbers. Consider the following tree T_n that has two kinds of internal nodes: AND-node and OR-node. The root of T_n is an OR-node with degree n. In general, an OR-node with degree $d \geq 2$ is called an **d-node**. If $d = 0$ or $d = 1$, then the d-node is simply a leaf (no children). For $d \geq 2$, each of the children of the d-node is an AND-node of degree exactly 2. Moreover, the ith child (for $i = 1, \ldots, d$) has two children which are an $(i-1)$-node and a $(n-i)$-node. This completes the description of T_n. Using T_n, we now define the sample space $S(T_n)$.

FIGURE

A sample point $\omega \in S(T_n)$ is a subtree of T_n, containing the following nodes: the root (which is the unique n-node of T_n) belongs to ω. In general, suppose $u \in \omega$. If u is an AND-node, then every child of u is in ω. If u is an OR-node, then exactly one child of u is in ω. This completes the description. What is the probability $\Pr(\omega)$? If ω has only one node, then $\Pr(\omega) = 1$. Otherwise, let ω_1, ω_2 be subtrees of ω, where the roots of ω_1, ω_2 are the grandchild of the root of ω. Then the probabilities $\Pr(\omega_1), \Pr(\omega_2)$ have been defined, and we have

$$\Pr(\omega) = \frac{1}{n} \Pr(\omega_1) \Pr(\omega_2).$$

This completely describes $S(T_n)$. There is another way to describe $S(T_n)$, as the set of all binary trees with exactly n nodes (internal or leaves). This is just a more compact way to encode the tree ω above.

Exercises

Exercise 1.1: Show that the method of assigning (uniform) probability to events in $B^1[a, b]$ is well-defined. ♢

Exercise 1.2: Let $\Omega = \mathbb{R}$. In the text, the event space defined Ω was restricted to a finite interval $[a, b]$. Define a probability space on Ω in which the entire real line is used in an essential way. ♢

Exercise 1.3: Consider the following randomized process, which is a sequence of steps. At each step, we roll a dice that has one of six possible outcomes: $1, 2, 3, 4, 5, 6$. In the i-th step, if the outcome is less than i, we stop. Otherwise, we go to the next step. The first step is $i = 1$. For instance, we never stop after first step, and surely stop by the 7-th step. Let T be the random variable corresponding to the number of steps.
(a) Set up the sample space, the event space, and the probability function for T.
(b) Compute the expected value of T. ♢

Exercise 1.4: J. Quick felt that the sample space S_n we constructed for Quicksort is unnecessarily complicated: why don't we define S_n to be the set of all permutations on the n input numbers. The probability of each permutation in S_n is $1/n!$. What is wrong with this suggestion? ♢

Exercise 1.5: Give simple upper and lower bounds on the size $C(n)$ of the sample space in Quicksort on n input numbers. Note that $C(n)$ are the Catalan numbers in Chapter 6 (see also Exercise there). ♢

Exercise 1.6: (Probabilistic Counters) Recall the counter problem where, given a binary counter C which is initially 0, you can perform the operation `inc`(C) to increments its value by 1. Now we want to do **probabilistic counting**: each time you call `inc`(C), it will flip a fair coin. If heads, the value of C is incremented and otherwise the value of C is unchanged. Now, at any moment you could call `look`(C), which will return *twice* the current value of C. Let X_m be the value of `look`(C) after you have made m calls to `inc`(C).
(a) Note that X_m is a random variable. What is the sample space Ω here?
(b) Let $P_m(i)$ be the probability that `look`$(C) = 2i$ after m `inc`'s. State a recurrence equation for $P_m(i)$ involving $P_{m-1}(i)$ and $P_{m-1}(i-1)$.
(c) Give the exact formula for $P_m(i)$ using binomial coefficients. HINT: you can either use the model in (a) to give a direct answer, or you can try to solve the recurrence of (b). You may recall that binomial identity $\binom{m}{i} = \binom{m-1}{i} + \binom{m-1}{i-1}$.
(d) In probabilistic counting we are interested in the *expected* value of `look`(C), namely $E[X_m]$. What is the expected value of X_m? HINT: express $E[X_m]$ using $P_m(i)$ and do some simple manipulation involving binomial coefficients. If you do not see what is coming out, try small examples like $m = 2, 3$ to see what the answer is.

Remarks: The expected value of X_m can be odd even when the actual value returned is always even. What have we gained by using this counter? We saved 1-bit! But, by a generalization of these ideas, you can probabilistically count to 2^{2^n} with an n-bit counter, thus saving exponentially many bits. ♢

Exercise 1.7: Let us prove the formula (**??**). To do this, consider the generating function

$$G(x) = \sum_{n=0}^{\infty} C(n)x^n.$$

Show that $G(x) = xG(x)^2 + 1$ and hence $G(x) = (1 - \sqrt{1-4x})/2x$. Use the Taylor expansion of $G(x)$ at $x = 0$. ♢

End Exercises

§2. Independence and Conditioning

Intuitively, the outcomes of two tosses of a coin ought to be "independent" of each other. In rolling a pair of dice, the probability of the event "the sum is at least 8" must surely be "conditioned by" the knowledge about the outcome of one of the dice. For instance, knowing that one of the dice is 1 or not critically affects this probability. We formalize such ideas of independence and conditioning.

Let $B \in \Sigma$ be any non-null event (*i.e.*, $\Pr(B) > 0$). Such an event B **induces** a probability space which we denote by $\Sigma|B$. The sample space of $\Sigma|B$ is B and event space is $\{A \cap B : A \in \Sigma\}$. The probability function $\Pr_B$ of the induced space is given by

$$\Pr_B(A \cap B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

It is conventional to write

$$\Pr(A|B)$$

instead of $\Pr_B(A \cap B)$, and to call it the **conditional probability of A given B**. Note that $\Pr(A|B)$ is undefined if $\Pr(B) = 0$.

Two events $A, B \in \Sigma$ are **independent** if $\Pr(AB) = \Pr(A)\Pr(B)$.

> Note that, for the first time, we have multiplied two probabilities! This is significant – in general whenever you multiply probabilities, there must be some independence requirement. Just as the product of two numbers x, y is usually written as xy with the $\times$ operator implicit, the intersection $A \cap B$ of two events is usually written AB. This analogy between intersection and multiplication is clarified through the concept of independence.
> Until now, we have only added probabilities, $\Pr(A)+\Pr(B)$. The conditions for adding probabilities are some disjointness requirement on events: $A \cap B = \emptyset$. The combination of adding and multiplying probabilities therefore brings a ring-like structure (involving $+, \times$) into play, and greatly enriches the subject.

It follows that if A, B are independent then $\Pr(A|B) = \Pr(A)$. More generally, a set $S \subseteq \Sigma$ of events is **k-wise independent** if for every subset $\{B_1, \ldots, B_m\} \subseteq S$ of m $(2 \leq m \leq k)$ distinct events, $\Pr(\cap_{i=1}^m B_i) = \prod_{i=1}^m \Pr(B_i)$. If $k = 2$, we say S is **pairwise independent**. If $k = |S|$, we simply say S is **independent**.

¶8. Bayes' Formula. Suppose $A_1, \ldots, A_n$ are mutually exclusive events such that $\Omega = \cup_{i=1}^n A_i$. Then for any event B, we have

$$\Pr(B) = \Pr(\uplus_{i=1}^n B \cap A_i) = \sum_{i=1}^n \Pr(B|A_i)\Pr(A_i). \tag{3}$$

Consider $\Pr(A_j|B) = \Pr(BA_j)/\Pr(B)$. If we replace the numerator by $\Pr(B|A_j)\Pr(A_j)$, and the denominator by (3), we obtain **Bayes' formula**,

$$\Pr(A_j|B) = \frac{\Pr(B|A_j)\Pr(A_j)}{\sum_{i=1}^n \Pr(B|A_i)\Pr(A_i)}. \tag{4}$$

In other words, this is a formula for inversion of conditional probability: given that you know B has occurred, you can determine the probability that any (mutually exclusive) A_j also occurred if you know $\Pr(B|A_i)$ for all i. This formula is the starting point for Bayesian probability, the empirical or predictive approach mentioned in the introduction. The goal of Bayesian probability is to use observations to predict the future.

¶9. Formula for Joint Events. From the definition of conditional probability, we have $\Pr(A_1A_2) = \Pr(A_1)\Pr(A_2|A_1)$, or more generally,

$$\Pr(A_1A_2|B) = \Pr(A_1|B)\Pr(A_2|A_1B).$$

This formula is generalized to: suppose $B, A_1, A_2, \ldots, A_n$ are events. Then

$$\Pr(\bigcap_{i=1}^{n} A_i|B) = \prod_{i=1}^{n} \Pr(A_i|A_1A_2\cdots A_{i-1}B). \tag{5}$$

In proof, simply expand the ith factor as $\Pr(A_1A_2, \ldots, A_{i-1}B)/\Pr(A_1A_2, \ldots, A_iB)$, and cancel common factors in the numerator and denominator. If $B = \Omega$, this reduces to

$$\Pr(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} \Pr(A_i|A_1A_2\cdots A_{i-1}).$$

E.g., $\Pr(ABCD) = \Pr(A)\Pr(B|A)\Pr(C|AB)\Pr(D|ABC)$. This formula is "extensible" in that the formula for $\Pr(A_1\cdots A_n)$ is derived from formula for $\Pr(A_1\cdots A_{n-1})$, by appending an extra factor.

Exercises

Exercise 2.1: Construct a set of events that is pairwise independent but not independent. HINT: Let $\Omega = \{1, 2, 3, 4\}$. Use the counting probability model for Ω, and consider the events $A = \{1, 2\}$, $B = \{1, 3\}$, $C = \{1, 4\}$. ♢

Exercise 2.2: In a popular TV game-show[2] called "Let's Make a Deal", there are three veiled stages. A prize car is placed behind one of these veils. Each contestant hopes to pick the stage with the car. The rules of the game are as follows: initially, the contestant picks one of the stages. Then the game-master selects one of the other two stages to be unveiled – this unveiled stage is inevitably car-less. The game-master now asks the contestent if he or she wishes to switch the original pick. There are two strategies to be analyzed: *always-switch* or *never-switch*. The never-switch strategy is easy to analyze: you have 1/3 chance of winning. Here are three conflicting claims about the always-switch strategy:
CLAIM I: your chance of winning is is 1/3, nothing has changed since the start.
CLAIM II: your chance of winning is 1/2, since the car is behind one of the two veiled stages.
CLAIM III: your chance of winning is 2/3, since it is the complement of the never-switch strategy.
(a) Find flaws in two of the claims.
(b) Set up a model to justify the unflawed claim. HINT: set up a sample space in which the sample points are paths in a tree and levels of the tree corresponds to various choices and decisions in the problem.
(c) Do we need the assumption that whenever the game-master has a choice of two stages to unveil, he picks either one with equal probability? ♢

Exercise 2.3: The above 2 strategies are deterministic. Actually, there is another reasonable strategy to examine. That is to flip a coin, and to switch only if it is heads. Analyze this randomized strategy. ♢

[2]This problem has generated some public interest, including angry letters by professional mathematicians to the New York Times claiming that there ought to be no difference in the two strategies described in the problem.

Exercise 2.4: Let us generalize the above game. The game begins with a car hidden behind one of $m \geq 4$ possible stages. After you make your choice, the game-master unveils all but two stages. Of course, the unveiled stages are all empty, and the two veiled stages always include one you picked.
(a) Analyze the always-switch strategy under the assumption that the game-master randomly picks the other stage.
(b) Suppose you want to assume the game-master is really trying to work against you. How does your analysis change? ♢

Exercise 2.5: The kind of probability space used in the above analysis is quite specialized in that it can be organized into a finite decision tree. The nodes at a given level $\ell \geq 0$ correspond to a decision variable x_ℓ. Each decision variable has a binary outcome (for simplicity). There are two players (0 and 1) corresponding who must make decisions at alternate levels. Player 0 (resp. player 1) correspond to the even (resp., odd) levels. There is a win/loss function $w(\sigma)$ that decides for each sequence of decisions whether player 1 wins. Suppose the game plan of player 0 is completely known (it can be probabilistic or deterministic). Does there always exist an optimal strategy for player 1? ♢

End Exercises

§3. Random Functions and Variables

The concepts so far have not risen much above the level of "gambling and parlor games" (the pedigree of our subject). Probability theory really takes off after we introduce the concept of random variables. Example of a random variable: using running example (E1), it is simply a function of the form $X : \Omega \to \mathbb{R}$ where $X(H) = 1$ and $X(T) = 0$. This random variable X has an expected (=average) value, namely, $E[X] = \Pr\{X = H\} \cdot X(H) + \Pr\{X = T\} \cdot X(T) = p \cdot 1 + (1-p) \cdot 0 = p$.

But random variables are just a special kind "random function". Let D be a set and $(\Omega, \Sigma, \Pr)$ a probability space. A **random function over** D is a function

$$f : \Omega \to D$$

such that for each $x \in D$, the set $f^{-1}(x)$ is an event. So that we may speak of the **probability** of x, *viz.*, $\Pr(f^{-1}(x))$. We also call $(\Omega, \Sigma, \Pr)$ the **underlying probability space** of f. We say f is **uniformly distributed on** D if $\Pr(f^{-1}(x)) = \Pr(f^{-1}(y))$ for all $x, y \in D$. We sometimes use bold fonts (**f** instead of f, etc) to denote random functions.

The most important random functions arise as follows: a **random variable** (r.v.) of a probability space $(\Omega, \Sigma, \Pr)$ is an real function

$$X : \Omega \to \mathbb{R}$$

such that for all $c \in \mathbb{R}$,

$$X^{-1}(H_c) = \{\omega \in \Omega \,:\, X(\omega) \leq c\}$$

belongs to Σ, where H_c is a generator of the Euclidean Borel field B^1. Sometimes the range of X is the extended reals $\mathbb{R} \cup \{\pm\infty\}$. It follows that for any Euclidean Borel set $A \in B^1$, the set

$$X^{-1}(A) = \{\omega \in \Omega \,:\, X(\omega) \in A\} \tag{6}$$

is an event. This event is usually written

$$\{X \in A\}. \tag{7}$$

In particular, $X^{-1}(c)$ is an event for all $c \in \mathbb{R}$, and so a r.v. is, *a fortiori*, a random object. In fact, a r.v. is just "a random real number".

Convention. Writing (7) for (6) illustrates the habit of probabilists to avoid explicitly mentioning sample points. More generally, probabilists will specify events by writing $\{\ldots X \ldots Y \ldots\}$ where "$\ldots X \ldots Y \ldots$" is some predicate on r.v.'s X, Y, etc. This really denotes the event $\{\omega \in \Omega : \ldots X(\omega) \ldots Y(\omega) \ldots\}$. For instance, $\{X \leq 5, X + Y > 3\}$ refers to the event $\{\omega \in \Omega : X(\omega) \leq 5, X(\omega) + Y(\omega) > 3\}$. Moreover, instead of writing $\Pr(\{\ldots\})$, we simply write $\Pr\{\ldots\}$, where the pair of curly brackets reminds that $\{\ldots\}$ is a set (which happens to be an event).

If X, Y are r.v.'s then so are

$$\min(X, Y), \quad \max(X, Y), \quad X + Y, \quad XY, \quad X^Y, \quad X/Y$$

where $Y \neq 0$ in the last case.

All random variables in probability theory are either discrete or continuous, which we now define. A r.v. X is **discrete** if the range of X is countable (this is automatic if Ω is countable). The special case[3] where the range is $\{0, 1\}$ is called a **Bernoulli r.v.**. We call X the **indicator function** of an event E if $X(\omega) = 1$ if $\omega \in E$ and $X(\omega) = 0$ else. Thus Bernoulli functions and indicator functions are basically synonymous.

A r.v. X is **continuous** if there exists a nonnegative function $f(x)$ defined for all $x \in \mathbb{R}$ such that for any Euclidean Borel set $A \in B^1$,

$$\Pr\{X \in A\} = \int_A f(x)dx$$

(cf. (7)). It follows that for any real $a \leq b$, $\Pr\{a \leq X \leq b\} = \int_a^b f(x)dx$ and hence $\Pr\{X = a\} = 0$. We call $f(x)$ the **density function** of X.

As examples of random variables, suppose in running example (E1), if we define $X(H) = 1, X(T) = 0$ then X is the indicator function of the "head event". For (E2), let us define $X(i) = i$ for all $i = 1, \ldots, 6$. If we have a game in which a player is paid i dollars whenever the player rolls an outcome of i, then X represents "payoff function".

¶10. Analysis of QuickSort. We re-visit the Quicksort algorithm from Lecture II.

Assume the input is an array $A[1..n]$ holding n numbers. Recall that Quicksort picks a random $r \in \{1, \ldots, n\}$ and uses the value $A[r]$ to partition the numbers in $A[1..n]$ into those that are greater than $A[r]$ and those that are less than $A[r]$. An elegant solution is possible where we do not use extra storage, and only move values within the array A. This is achieved by the following Partition subroutine:

[3] In another variant, the range is $\{+1, -1\}$ and is used in discrepancy theory.

```
Partition(A, i, j, e):
Input: Array A[1..n] and 1 ≤ i < j ≤ n, with e occurring in A[i..j]
Output: Index k ∈ {i, ..., j} the subarray A[i..j] is rearranged
so that A[k] = e, A[i..k − 1] ≤ e, and A[k + 1..j] ≥ e.
    while j − i ≥ 1
        while (i < j and A[++i] ≤ e);
        while (i < j and A[−−j] ≥ e);
        Swap A[i] ↔ A[j]
    return(j)
```

To sort the array $A[1..n]$, we invoke QuickSort$(A, 1, n)$, where QuickSort is the following recursive procedure:

```
QuickSort(A, i, j):
Input: Array A[1..n] and 1 ≤ i < j ≤ n.
Output: The subarray A[i..j] is sorted in non-decreasing order.
    if i = j, return(A).
    Randomly pick a r ∈ {i, ..., j};
    k ←Partition(A, 1, n, A[r])
    QuickSort(A, i, k)
    QuickSort(A, k, j)
```

Let us simplify the analysis by assuming the input numbers are distinct. Suppose the set of numbers in $A[1..n]$ is

$$Z := \{z_1, \ldots, z_n\}$$

where $z_1 < z_2 < \cdots < z_n$. For each $1 \le i < j \le n$, let

$$E_{ij} = \{z_i \text{is compared to } z_j\}$$

denote the event that that z_i and z_j are compared. Let X_{ij} be the indicator function for E_{ij}. If X is the random variable for the number of comparisons in quicksort, then we have

$$X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij}.$$

The critical observation is

Lemma 1. $\Pr(E_{ij}) = \frac{2}{j-i+1}$

From this lemma, we see that

$$\mathtt{E}[X] < \sum_{i=1}^{n-1} 2H_n < 2n \lg n.$$

It remains to prove Lemma 1. We will prove a more general statement: let A_{ij} be the event that the pivot r on input $Z = \{1, \ldots, n\}$ satisfies the property $r < i$ or $r > j$. For all $1 \le i < j \le n$, let E^n_{ij} denote the event that the comparison $z_i : z_j$ occurred on input $Z = \{1, \ldots, n\}$. When $i = 1$ and $j = n$, we let

$$E^n := E^n_{1,n}.$$

We claim: if $1 < i$ or $j < n$ then

$$\Pr(E^n_{ij}|A_{ij}) = \Pr(E^{j-i+1}). \tag{8}$$

Since $1 < i$ or $j < n$, we must have $n - j + i - 1 \geq 1$. We use induction on $n - j + i - 1$. The result is clearly true when $n - j + i - 1 = 1$. Otherwise, we see that

$$\begin{aligned}
\Pr(E^n_{ij}|A_{ij}) \cdot \Pr(A_{ij}) &= \Pr(E^n_{ij} A_{ij}) \\
&= \Pr(E^n_{ij}|q < i) \Pr\{q < i\} + \Pr(E^n_{ij}|q > j) \Pr\{q > j\} \\
&= \Pr(E^{j-i+1}) \Pr\{q < i\} + \Pr(E^{j-i+1}) \Pr\{q > j\} \qquad \text{(by induction hypothesis)} \\
&= \Pr(E^{j-i+1}) \left[\tfrac{i-1}{n} + \tfrac{n-j}{n}\right] \\
&= \Pr(E^{j-i+1}) \Pr(A_{ij}).
\end{aligned}$$

But it is immediate that

$$\Pr(E^n) = \frac{2}{n}$$

and thus $\Pr(E^n_{ij}) = \Pr(E^{j-i+1}) = \frac{2}{j-i+1}$. This proves Lemma 1.

¶11. Random Objects. If the elements of D are objects of some category t of objects, we may also call the random function $f : \Omega \to D$ a **random t object**. Examples:

- If D is some set of graphs we call f a **random graph**.
- For any set S, we call f a **random k-set** of S if $D = \binom{S}{k}$. If D is the set of permutations of S, then f is a **random permutation** of S.
- More generally, if D is some arbitrary set, we may call f a **random D-element**.

Discussion: The power of random objects is that they are composites of the individual objects of D. For all many purposes, these objects are as good as the honest-to-goodness objects in D. Another view of this phenomenon is to use the philosophical idea of alternative or possible worlds. Each $\omega \in \Omega$ is a possible world[4] Then $f(\omega)$ is just the particular incarnation of f in the world ω.

Example: (Finite Field Space) Consider the uniform probability space on $\Omega = F^2$ where F is any finite field. For each $x \in F$, consider the random function

$$\begin{aligned}
&\mathbf{h}_x : \Omega \to F, \\
&\mathbf{h}_x(\langle a, b\rangle) = ax + b, \qquad (\langle a, b\rangle \in \Omega).
\end{aligned}$$

We claim that $\mathbf{h}_x$ is a random element of F, *i.e.*, $\Pr\{\mathbf{h}_x = i\} = 1/|F|$ for each $i \in F$, This amounts to saying that there are exactly $|F|$ sample points $\langle a, b\rangle = \omega$ such that $\mathbf{h}_x(\omega) = i$. To see this, consider two cases: (1) If $x = 0$ then clearly $b = i$ and a can be arbitrarily chosen. (2) If $x \neq 0$, then for any choice of b, there is unique choice of a, namely $a = (i - b)x^{-1}$.

Example: (Random Graphs) Fix $0 \leq p \leq 1$ and $n \geq 2$. Consider the probability space where $\Omega = \{0, 1\}^m$, $m = \binom{n}{2}$, $\Sigma = 2^\Omega$ and for $(b_1, \ldots, b_m) \in \Omega$, $\Pr(b_1, \ldots, b_m) = p^k(1 - p)^{m-k}$ where k is the number of 1's in $(b_1, \ldots, b_m)$. Once checks

[4]Good thing too, ω can be confused with the letter w.

that Pr as defined is a probability function. Let K_n be the complete bigraph on n vertices whose edges are labelled with the integers $1, \ldots, m$. Consider the random graph

$$G_{n,p} : \Omega \to \text{subgraphs of } K_n \tag{9}$$

where $G_{n,p}(b_1, \ldots, b_m)$ is the subgraph of K_n with precisely those edges that are labeled i where $b_i = 1$.

¶12. Random Statistics. Random variables often arise as follows. A function $C : D \to \mathbb{R}$ is called a **statistic** of D where D is some set of objects. If $g : \Omega \to D$ is a random object, we obtain the random variable $C_g : \Omega \to \mathbb{R}$ where

$$C_g(\omega) = C(g(\omega)).$$

Call C_g a **random statistic** of g.

For example, let $g = G_{n,p}$ be the random graph in equation (9). and let C count the number of Hamiltonian cycles in a bigraph. Then the random variable

$$C_g : \Omega \to \mathbb{R} \tag{10}$$

is defined so that $C_g(\omega)$ is the number of Hamiltonian cycles in $g(\omega)$.

¶13. k-Wise Independence. We extend some concepts of independence from events to random variables.

A collection $\{X_1, X_2, \ldots, X_n\}$ of n r.v.'s is **k-wise independent** (some $k \geq 2$) if for all $c_1, \ldots, c_n \in \mathbb{R}$, the events $\{X_1 \leq c_1\}, \ldots, \{X_n \leq c_n\}$ are k-wise independent. If $k = 2$, we say K is **pairwise independent**. The collection K is **independent** if if is k-wise independent for all $k = 2, \ldots, n$. An infinite collection of r.v.'s is **(k-wise) independent** if every finite subcollection is (k-wise) independent.

Let D be a set. A set $K = \{f_1, \ldots, f_n\}$ of random D-objects is called an **ensemble** if the f_i's have a common underlying probability space. If D is finite, we say K is **k-wise independent** if for any $a_1, \ldots, a_k \in D$, $\Pr\{f_1 = a_1, \ldots, f_k = a_k\} = \prod_{i=1}^{n} \Pr\{f_i = a_i\}$.

Example: (Finite Field Space) Recall the finite field space $\Omega = F^2$ above. Let

$$K = \{\mathbf{h}_x : x \in F\} \tag{11}$$

where $\mathbf{h}_x(\langle a, b\rangle) = ax + b$ as before. We have shown that each $\mathbf{h}_x$ is a random element of F. We now claim that the elements in K are pairwise independent. Fix $x, y, i, j \in F$ and let $n = |F|$. Suppose $x \neq y$ and $\mathbf{h}_x = i$ and $\mathbf{h}_y = j$. This means

$$\begin{pmatrix} x & 1 \\ y & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} i \\ j \end{pmatrix}.$$

The 2×2 matrix is invertible and hence (a, b) has a unique solution. Hence

$$\Pr\{\mathbf{h}_x = i, \mathbf{h}_y = j\} = 1/n^2 = \Pr\{\mathbf{h}_x = i\} \Pr\{\mathbf{h}_y = j\},$$

as desired.

Algorithmically, constructions of k-wise independent variables over an underlying probability space that is small (in this case, $|\Omega| = p^2$) is important because it allows us to make certain probabilistic constructions effective.

Exercises

Exercise 3.1: Compute the probability of the event $\{C_g = 0\}$ where C_g is given by (10). Do this for $n = 2, 3, 4$. ♢

Exercise 3.2: Consider the following (silly) randomized process, which is a sequence of probabilistic steps. At each step, we roll a dice that has one of six possible outcomes: $1, 2, 3, 4, 5, 6$. In the i-th step, if the outcome is less than i, we stop. Otherwise, we go to the next step. The first step is $i = 1$. For instance, we never stop after first step, and surely stop by the 7-th step. Let T be the random variable corresponding to the number of steps.
(a) Set up the sample space, the event space, and the probability function for T.
(b) Compute the expected value of T. ♢

Exercise 3.3: Let U be a finite set, $|U| = n$, and $\Pi(U)$ the set of permutations of U. Let $S : \Omega \to 2^U$ be a random subset of U and

$$P : \Omega \to \bigcup_{V \subseteq U} \Pi(V).$$

We say P is a **permutation** of S if $P(\omega) \in \Pi(S(\omega))$ for all $\omega \in \Omega$. If, for each subset $V \subseteq U$ and $\pi \in \Pi(V)$, $\Pr\{P = \pi | S = V\} = 1/(m!)$ where $m = |V|$, then we call P a **uniform random permutation** of S. Explicitly construct a probability space Ω and random functions P, S such that P is a uniform random permutation of S. ♢

Exercise 3.4: Let K be the set of random elements in the finite field F given by (11).
(a) Show that K is not 3-wise independent.
(b) Generalize the example to construct a collection of k-wise independent random functions. ♢

Exercise 3.5: Let $W(n, x)$ (where $n \in \mathbb{N}$ and $x \in \mathbb{Z}_n$) be a "witness" predicate for compositeness: if n is composite, then $W(n, x) = 1$ for at least $n/2$ choices of x; if n is prime, then $W(n, x) = 0$ for all x. Let $W(n)$ be the random variable whose value is determined by a random choice of x. Let $W_t(n)$ be the random variable whose value is obtained as follows: randomly choose n values $x_1, \ldots, x_n \in \mathbb{Z}_n$ and compute each $W(n, x_i)$. If any $W(n, x_i) = 1$ then $W_t(n) = 1$ but otherwise $W_t(n) = 0$.
(a) If n is composite, what is the probability that $W_t(n) = 1$?
(b) Now we compute $W_t(n)$ using somewhat less randomness: first assume t is prime and larger than n. only randomly choose two values $a, b \in \mathbb{Z}_t$. Then we define $y_i = a \cdot i + b \pmod t$. We evaluate $W_t(n)$ as before, except that we use $y_0, \ldots, y_{t-1} \pmod n$ instead of the x_i's. Lower bound the probability that $W_t(n) = 1$ in this new setting. ♢

Exercise 3.6:

(a) If a collection of r.v.'s is k-wise independent, then it is also $(k-1)$-wise independent.

(b) Let $f_1, \ldots, f_n$ be real functions $f_i : \mathbb{R} \to \mathbb{R}$ and $\{X_1, \ldots, X_n\}$ is a set of independent r.v.'s. If $f_i(X_i)$ are also r.v.'s then $\{f_1(X_1), \ldots, f_n(X_n)\}$ is also a set of independent r.v.'s. ♢

§4. Random Number Generation and Applications

Without question, the most important primitive in any computational model that supports randomized algorithms is the **random number generator**. This is a function which, when called with no arguments, returns a real number in the unit interval $[0, 1]$. This defines a random variable $U_{[0,1]}$ which is uniformly distributed over the unit interval. This is a purely theoretical construct. In practice, some discrete approximation to $U_{[0,1]}$ is used.

In most programming languages, or at least in the standard libraries for the language, there is a function called `random()` (or perhaps `rand()`) which returns a machine representable number in the half-open interval $[0, 1)$, and whose distribution is a good approximation to the uniform distribution. Although our main interest in random number generators is mainly in the context of randomized algorithms, it has remarkably many other applications: simulation of natural phenomena (computer graphics effects, weather, etc), testing of systems for defects, sampling of populations, decision making and in recreation (dice, card games, etc).

We want to address another basic primitive: random permutations. Fix a natural number $n \geq 2$. Let S_n denote the set of permutations on $[1..n]$. A **random permutation** P of S_n is just a random function p such that $\Pr\{p = \pi\} = 1/n!$ for all $\pi \in S_n$. We may choose $(\Omega, \Sigma) = (S_n, 2^\Omega)$ as the underlying event space.

Our problem is that of constructing P starting from a random number generator. Here is an extremely simple algorithm from Moses and Oakford (see [5, p. 139]).

```
RandomPermutation
    Input: an array A[1..n].
    Output: A random permutation of S_n stored in A[1..n].
1.  for i = 1 to n do        // Initialize array A
2.      A[i] = i.
3.  for i = n downto 2 do    // Main Loop
4.      X ← 1 + ⌊i · random()⌋.
5.      Exchange contents of A[i] and A[X].
```

This algorithm takes linear time; it makes $n-1$ calls to the random number generator and makes $n-1$ exchanges of a pair of contents in the array. Here is the correctness assertion for this algorithm:

Lemma 2. *Every permutation of* $[1..n]$ *is equally likely to be generated.*

Proof. The proof is as simple as the algorithm. Pick any permutation σ of $[1..n]$. Let A' be the value of the array A at the end of running this algorithm. So it is enough to prove that

$$\Pr(A' = \sigma) = \frac{1}{n!}.$$

Let E_i be the event $\{A'[i] = \sigma(i)\}$, for $i = 1, \ldots, n$. Thus

$$\Pr(A' = \sigma) = \Pr(E_1 E_2 E_3 \cdots E_{n-1} E_n).$$

First, note that $\Pr(E_n) = 1/n$. Also, $\Pr(E_{n-1}|E_n) = 1/(n-1)$. In general, we see that

$$\Pr(E_i | E_n E_{n-1} \cdots E_{i+1}) = \frac{1}{i}.$$

The lemma now follows from an application of (5) which shows $\Pr(E_1 E_2 E_3 \cdots E_{n-1} E_n) = 1/n!$. **Q.E.D.**

Note that the conclusion of the lemma holds even if we initialize the array A with any permutation of $[1..n]$. This fact is useful if we need to computer another random permutation in the same array A.

While the above analysis is simple, it is instructive to ask what is the underlying probabily space? Basically, if A' is the value of the array at the end of the algorithm, then A' is a random permutation in the sense of §3. That is,

$$A' : \Omega \to S_n$$

where Ω is a suitable probability space and S_n is the set of n-permutations. We can view Ω as the set $\prod_{i=2}^{n}[0,1)$ where a typical $\omega \in \Omega = (x_2, x_3, \ldots, x_n)$ tells us the sequence of values returned by the $n-1$ calls to the `random()` function.

Remarks: Random number generation is an extensively studied topic: Knuth [5] is a basic reference. The concept of randomness is by no means easily pinned down. From the complexity viewpoint, there is a very fruitful approach to randomness called Kolmogorov Complexity. A comprehensive treatment is found in Li and Vitányi [7].

§5. Expectation and Variance

Two important numbers are associated with a random variable: its "average value" and its "variance" (likelihood of deviating from the average value).

If X is a discrete r.v. whose range is

$$\{a_1, a_2, a_3 \ldots\} \tag{12}$$

then its **expectation** (or, **mean**) $\mathtt{E}[X]$ is defined to be

$$\mathtt{E}[X] := \sum_{i \geq 1} a_i \Pr\{X = a_i\}.$$

This is well-defined provided the series converges absolutely, *i.e.*, $\sum_{i\geq 1} |a_1| \Pr\{X = a_i\}$ converges. If X is a continuous r.v. with probability density $f(x)$ then

$$\mathtt{E}[X] := \int_{-\infty}^{\infty} u f(u) du.$$

Note that if X is the indicator variable for an event A then

$$\mathtt{E}[X] = \Pr(A).$$

¶14. Two Remarkable Properties of Expectation. The following two elementary properties of expectation can often yield surprising consequences.

The first property is the **linearity** of expectation. This means that for all r.v.'s X, Y and $\alpha, \beta \in \mathbb{R}$:

$$\mathtt{E}[\alpha X + \beta Y] = \alpha \mathtt{E}[X] + \beta \mathtt{E}[Y].$$

The remarkable fact is that X and Y are completely arbitrary – for instance, we need no independence assumptions. In applications, we can often decompose a r.v. X into *any* linear combination of r.v.s $X_1, X_2, \ldots, X_m$. If we can compute the expections of each X_i, then by linearity of expectation, we obtain the expection of X itself. Typically, X may be the running time of a n-step algorithm and X_i is the expected time for the ith step.

The second property is that, from expectations, we can assert the existence of objects with certain properties.

Lemma 3. *Suppose X is a discrete r. v. with finite expectation μ. If Ω is finite, then:*

(i) *There exists $\omega_0, \omega_1 \in \Omega$ such that*

$$X(\omega_0) \leq \mu \leq X(\omega_1). \tag{13}$$

(ii) *If X is non-negative, then*

$$\Pr\{X \leq 2\mu\} \geq 1/2. \tag{14}$$

In particular, if $\Pr\{\cdot\}$ is uniform and Ω finite, then at least half of the sample points $\omega \in \Omega$ satisfy $X(\omega) \leq 2\mu$.

Proof. Since X is discrete, let

$$\mu = \mathtt{E}[X] = \sum_{i=1}^{\infty} a_i \Pr\{X = a_i\}.$$

(i) If there are arbitrarily negative a_i's then clearly ω_0 exists; otherwise choose ω_0 so that $X(\omega_0) = \inf\{X(\omega) : \omega \in \Omega\}$. Likewise if there are arbitrarily large a_i's then ω_1 exists, and otherwise choose ω_1 so that $X(\omega_1) = \sup\{X(\omega) : \omega \in \Omega\}$. In every case, we have chosen ω_0 and ω_1 so that the following inequality confirms our lemma:

$$X(\omega_0) = X(\omega_0) \sum_{\omega \in \Omega} \Pr(\omega) \leq \sum_{\omega \in \Omega} \Pr(\omega) X(\omega) \leq X(\omega_1) \sum_{\omega \in \Omega} \Pr(\omega) = X(\omega_1).$$

(ii) This is just Markov's inequality. **Q.E.D.**

Let us apply this lemma to assert the existence of certain objects. Suppose we set up a random D object,

$$g : \Omega \to D$$

and are interested in a certain statistic $C : D \to \mathbb{R}$. Define the random statistic $C_g : \Omega \to \mathbb{R}$ as in (10). Then there exists ω_0 such that

$$C_g(\omega_0) \leq \mathtt{E}[C_g]$$

This means that the object $g(\omega_0) \in D$ has the property $C(g(\omega_0)) \leq \mathtt{E}[C_g]$.

Linearity of expectation amounts to saying that summing r.v.'s is commutative with taking expectation. What about products of r.v.'s? If X, Y are independent then

$$\mathtt{E}[XY] = \mathtt{E}[X]\mathtt{E}[Y]. \tag{15}$$

The requirement that X, Y be independent is necessary. As noted earlier, all multiplicative properties of probability depends from some form of independence.

The jth **moment** of X is $\mathtt{E}[X^j]$. If $\mathtt{E}[X]$ is finite, then we define the **variance** of X to be

$$\mathtt{Var}(X) := \mathtt{E}[(X - \mathtt{E}[X])^2].$$

Note that $X - \mathtt{E}[X]$ is the deviation of X from its mean. It is easy to see that

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

The positive square-root of $\mathtt{Var}(X)$ is called its **standard deviation** and denoted $\sigma(X)$ (so $\mathtt{Var}(X)$ is also written $\sigma^2(X)$). If X, Y are independent, then summing r.v.'s also commutes with taking variances. More generally:

Lemma 4. *Let X_i $(i = 1, \ldots, n)$ be pairwise independent random variables with finite variances. Then*

$$\mathtt{Var}(\sum_{i}^{n} X_i) = \sum_{i=1}^{n} \mathtt{Var}(X_i).$$

This is a straightforward computation, using the fact that $\mathtt{E}[X_i X_j] = \mathtt{E}[X_i]\mathtt{E}[X_j]$ for $i \neq j$ since X_i and X_j are independent.

¶15. Distribution and Density. For any r.v. X, we define its **distribution function** to be $F_X : \mathbb{R} \to [0, 1]$ where

$$F_X(c) := \Pr\{X \leq c\}, \qquad c \in \mathbb{R}.$$

The importance of distribution functions stems from the fact that the basic properties of random variables can be studied from their distribution function alone.

Two r.v.'s X, Y can be related as follows: we say X **stochastically dominates** Y, written

$$X \succeq Y$$

if $F_X(c) \leq F_Y(c)$ for all c. It is not hard to see (Exercise) that this implies $\mathtt{E}[X] \geq \mathtt{E}[Y]$ if X stochastically dominates Y. If $X \succeq Y$ and $Y \succeq X$ then we say they are **identically distributed**, denoted

$$X \sim Y.$$

A common probabilistic setting is a collection K of r.v.'s that is independent and with all the r.v.'s in K sharing the same distribution. We then say K is **independent**

and identically distributed (abbrev. i.i.d). For instance, when X_i is the outcome of the ith toss of some fixed coin, then $K = \{X_i\}$ is an i.i.d. family.

In general, a distribution function[5] $F(x)$ is a monotone non-decreasing real function such that $F(-\infty) = 0$ and $F(+\infty) = 1$. Sometimes, a distribution function $F(x)$ is defined via a **density function** $f(u) \geq 0$, where

$$F(x) = \int_{-\infty}^{x} f(u)du.$$

In case X is discrete, the density function $f_X(u)$ (of its distribution function F_X) is zero at all but countably many values of u. As defined above, a continuous r.v. X is specified by its density function.

¶16. Conditional Expectation. This concept is useful for computing expectation. if A is an event, define the **conditional expectation** $\mathtt{E}[X|A]$ of X to be $\sum_{i\geq 1} a_i \Pr\{X = a_i|A\}$. In the discrete event space, we get

$$\mathtt{E}[X|A] = \frac{\sum_{\omega\in A} X(\omega)\Pr(\omega)}{\Pr(A)}.$$

If B is the complement of A, then

$$\mathtt{E}[X] = \mathtt{E}[X|A]\Pr(A) + \mathtt{E}[X|B]\Pr(B).$$

More generally, if Y is another r.v., we define a new r.v. $Z = \mathtt{E}[X|Y]$ where $Z(\omega) = \mathtt{E}[X|Y = Y(\omega)]$ for any $\omega \in \Omega$. Thus $Z(\omega)$ depends only on $Y(\omega)$. We can compute the expectation of X using the formula

$$\begin{aligned} \mathtt{E}[X] &= \mathtt{E}[\mathtt{E}[X|Y]] && (16)\\ &= \sum_{a\in\mathbb{R}} \mathtt{E}[X|Y = a]\Pr\{Y = a\}. && (17) \end{aligned}$$

For example, let X_i's be i.i.d., and N be a non-negative integer r.v. independent of the X_i's. What is the expected value of $\sum_{i=1}^{N} X_i$?

$$\begin{aligned} \mathtt{E}[\sum_{i=1}^{N} X_i] &= \mathtt{E}[\mathtt{E}[\sum_{i=1}^{N} X_i|N]]\\ &= \sum_{n\in\mathbb{N}} \mathtt{E}[\sum_{i=1}^{N} X_i|N = n]\Pr\{N = n\}\\ &= \sum_{n\in\mathbb{N}} n\mathtt{E}[X_1]\Pr\{N = n\}\\ &= \mathtt{E}[X_1]\mathtt{E}[N]. \end{aligned}$$

We can also use conditioning in computing variance, since $\mathtt{E}[X^2)] = \mathtt{E}[\mathtt{E}[X^2|Y]]$.

Exercises

[5]Some authors calls the function $\Pr : \Sigma \to [0, 1]$ a "(probability) distribution" on the set Ω. We avoid this terminology.

Exercise 5.1: Answer YES or NO to the following question. A correct answer is worth 5 points, but a wrong answer gets you -3 points. Of course, if you do not answer, you get 0 points. "In a True/False question, you get 5 points for correct answer, 0 points for not attempting the question and -3 points for an incorrect guess. Suppose have NO idea what the answer might be. Should you attempt to answer the question?". ♢

Exercise 5.2: You face a multiple-choice question with 4 possible choices. If you answer the question, you get 6 points if correct and -3 if wrong. If you do not attempt the question, you get -1 point. Should you attempt to answer the question if you have no clue as to what the question is about? You must justify your answer to receive any credit. NOTE: *this* is not a multiple choice question. ♢

Exercise 5.3: You (as someone who is designing an examination) wants to assign points to a multiple choice question in which the student must pick one out of 5 possible choices. The student is not allowed to ignore the question. How do you assign points so that (a) if a student has no clue, then the expected score is -1 points and (b) if a student could eliminate one out of the 5 choices, the expected score is 0 points. ♢

Exercise 5.4: Compute the expected value of the r.v. C_g in equation (10) for small values of n ($n = 2, 3, 4, 5$). ♢

Exercise 5.5: Simple dice game: you are charged c dollars for rolling a dice, and if your roll has outcome i, you win i dollars. What is the fair value of c? HINT: what is your expected win per roll? ♢

Exercise 5.6: (a) Professor Vegas introduces a game of dice in class (strictly for "object lesson" of course). Anyone in class can play. To play the game, you pay \$12 and roll a pair of dice. If the product of the rolled values on the dice is n, then Professor Vegas pays you \$ n. For instance, if you rolled the numbers 5 and 6 then you make a profit of $\$18 = 30 - 12$. Student Smart would not play, claiming: *the probability of losing money is more than the probability of winning money.*

(a) What is right and wrong with Student Smart's claim?

(b) Would you play this game? Justify. ♢

Exercise 5.7: One day, Professor Vegas forgot to bring his pair of dice. He stills wants to play the game in the previous exercise. Professor Vegas decides to simulate the dice by tossing a fair coin 6 times. Interpreting heads as 1 and tails as zero, this gives 6 bits which can be viewed as two binary numbers $x = x_2x_1x_0$ and $y = y_2y_1y_0$. So x and y are between 0 and 7. If x or y is either 0 or 7 then the Professor returns your \$12 (the game is off). Otherwise, this is like the dice game in (a). What is the expected profit of this game? ♢

Exercise 5.8: In the previous question, we "simulate" rolling a dice by tossing three fair coins. Unfortunately, if the value of the tosses is 0 or 7, we

call off the game. Now, we want to continue tossing coins until we get a value between 1 and 6.
(a) An obvious strategy is this: each time you get 0 or 7, you toss another three coins. This is repeated as many times as needed. What is the expected number of coin tosses to "simulate" a dice roll using this method?
(b) Modify the above strategy to simulate a dice roll with fewer coin tosses. You need to (i) justify that your new strategy simulates a fair dice and (ii) compute the expected number of coin tosses.
(c) Can you show what the the optimum strategy is? ♢

Exercise 5.9: In the dice game of the previous exercise, Student Smart decided to do another computation. He sets up a sample space

$$S = \{11, 12, \ldots, 16, 22, 23, \ldots, 26, 33, \ldots, 36, 44, 45, 46, 55, 56, 66\}.$$

So $|S| = 21$. Then he defines the r.v. X where $X(ij) = i \times j$ and computes the expectation of X where using $\Pr(ij) = 1/21$. What is wrong? Can you correct his mistake without changing his choice of sample space? What is the alternative sample space? In what sense is Smart's choice of S is better? ♢

Exercise 5.10: Prove if $X \succeq Y$ then $\mathtt{E}[X] \geq \mathtt{E}[Y]$. Moreover, equality holds iff $X \sim Y$. ♢

Exercise 5.11: (a) Show that in any graph with n vertices and e edges, there exists a bipartite subgraph with $e/2$ edges. In addition, the bipartite subgraph have $\lfloor n \rfloor$ vertices on one side and $\lceil n \rceil$ of the other. Remark: depending on your approach, you may not be able to fulfil the additional requirement.
(b) Obtain the same result constructively (*i.e.*, give a randomized algorithm). ♢

Exercise 5.12: (Cauchy-Schwartz Inequality) Show that $\mathtt{E}[XY]^2 \leq \mathtt{E}[X^2]\mathtt{E}[Y^2]$ assuming X, Y have finite variances. ♢

Exercise 5.13: (Law of Unconscious Statistician) If X is a discrete r.v. with probability mass function $f_X(u)$, and g is a real function then

$$\mathtt{E}[g(X)] = \sum_{u: f_X(u) > 0} g(u) f_X(u).$$

♢

Exercise 5.14: If $X_1, X_2, \ldots$ are i.i.d. and $N \geq 0$ is an independent r.v. that is integer-valued then $\mathtt{E}[\sum_{i=1}^{N} X_i] = \mathtt{E}[N]\mathtt{E}[X_1]$ and $\mathtt{Var}(\sum_{i=1}^{N} X_i) = \mathtt{E}[N]\mathtt{Var}(X_1) + \mathtt{E}[X]^2\mathtt{Var}(N)$. ♢

Exercise 5.15: Suppose we have a fair game in which you can bet any dollar amount. If you bet \$x, and you win, you receive \$x; and otherwise you lose \$x.
(a) A well-known "gambling technique" is to begin by betting \$1. Each time

you lose, you double the amount of the bet (to \$2, \$4, etc). You stop at the first time you win. What is wrong with this scenario?
(b) Suppose you have a limited amount of dollars, and you want to devise a strategy in which the *probability* of your winning is as big as possible. (We are not talking about your "expected win".) How would you achieve this? ♢

Exercise 5.16: [Amer. Math. Monthly] A set consisting of n men and n women are partitioned at random into n disjoint pairs of people. Let X be the number of male-female couples that result. What is the expected value and variance of X? HINT: let X_i be the indicator variable for the event that the ith man is paired with a woman. To compute the variance, first compute $\mathtt{E}[X_i^2]$ and $\mathtt{E}[X_i X_j]$ for $i \neq j$. ♢

Exercise 5.17: [Mean and Variance of a geometric distribution] Let X be the number of coin tosses needed until the first head appears. Assume the probability of coming up heads is p. Use conditional probability (16) to compute $\mathtt{E}[X]$ and $\mathtt{Var}(X)$. HINT: let $Y = 1$ if the first toss is a head, and $Y = 0$ else. ♢

§6. Families of Random Variables

We now consider families of random variables over a common probability space. Two common situations arise.

(i) Perhaps the most important situation is when a family K of r.v.'s is i.i.d.

(ii) Another situation is when we have a family $\{X_t : t \in T\}$ of r.v.'s where $T \subseteq \mathbb{R}$ is the index set. We think of T as time and X_t as describing the behavior of a stochastic phenomenon evolving over time. Such a family is called a **stochastic process**. Usually $T = \mathbb{R}$ (continuous time) or $T = \mathbb{N}$ (discrete time).

We state two results that lay claim to being the fundamental theorems of probability theory. Both relate to i.i.d. families. Let $X_1, X_2, X_3, \ldots$, be a countable i.i.d. family of Bernoulli r.v.'s. Let $S_n := \sum_{i=1}^n X_i$ and $pas\, \Pr\{X_1 = 1\}$. It is intuitively clear that S_n approaches np as $n \to \infty$.

Theorem 5 ((Strong) Law of Large Numbers). *For any $\varepsilon > 0$, with probability 1, there are only finitely many sample points in the event*

$$|S_n - np| > \varepsilon$$

Theorem 6 (Central Limit Theorem). *See Ross*

¶17. Some probability distributions. The above theorems do not make any assumptions about the underlying distributions of the r.v.'s (therein lies their power). However, certain probability distributions are quite common and it is important to recognize them. Below we list some of them. In each case, we only need to describe the corresponding density functions $f(u)$. In the discrete case, it suffices to specify $f(u)$ at those elementary events u where $f(u) > 0$.

- **Binomial distribution** $B(n, p)$, with parameters $n \geq 1$ and $0 < p < 1$:

$$f(i) = \binom{n}{i} p^i (1-p)^{n-i}, \qquad (i = 0, 1, \ldots, n).$$

Sometimes $f(i)$ is also written $B_i(n,p)$ and corresponds to the probability of i successes out of n Bernoulli trials. In case $n = 1$, this is also called the Bernoulli distribution. If X has such a distribution, then

$$\mathtt{E}[X] = np, \quad \mathtt{Var}(X) = npq$$

where $q = 1 - p$.

- **Geometric distribution** with parameter p, $0 < p < 1$:

$$f(i) = p(1-p)^{i-1} = pq^{i-1}, \qquad (i = 1, 2, \ldots).$$

Thus $f(i)$ may be interpreted as the probability of the first success occurring at the ith Bernoulli trial. If X has such a distribution, then $\mathtt{E}[X] = 1/p$ and $\mathtt{Var}(X) = q/p^2$.

- **Poisson distribution** with parameter $\lambda > 0$:

$$f(i) = e^{-\lambda}\frac{\lambda^i}{i!}, \qquad (i = 0, 1, \ldots).$$

We may view $f(i)$ as the limiting case of $B_i(n,p)$ where $n \to \infty$ and $np = \lambda$. If X has such a distribution, then $\mathtt{E}[X] = \mathtt{Var}(X) = \lambda$.

- **Uniform distribution** over the real interval $[a, b]$:

$$f(u) = \begin{cases} \frac{1}{b-a} & a < u < b \\ 0 & \text{else.} \end{cases}$$

- **Exponential distribution** with parameter $\lambda > 0$:

$$f(u) = \begin{cases} \lambda e^{-\lambda u} & u \geq 0 \\ 0 & \text{else.} \end{cases}$$

- **Normal distribution** with mean μ and variance σ^2:

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right].$$

In case $\mu = 0$ and $\sigma^2 = 1$, we call this the unit normal distribution.

Exercises

Exercise 6.1: Verify the values of $\mathtt{E}[X]$ and $\mathtt{Var}(X)$ asserted for the various distributions of X.

♢

Exercise 6.2: Show that the density functions $f(u)$ above truly define distribution functions: $f(u) \geq 0$ and $\int_{-\infty}^{\infty} f(u)du = 1$. Determine the distribution function in each case. ♢

§7. Estimates and Inequalities

A fundamental skill in probabilistic analysis is estimating probabilities because they are often too intricate to determine exactly. We list some useful inequalities and estimation techniques.

¶18. Approximating the binomial coefficients. Recall Stirling's approximation in Lecture II.2. Using such bounds, we can show [8] that for $0 < p < 1$ and $q = 1 - p$,

$$G(p,n)e^{-\frac{1}{12pn}-\frac{1}{12qn}} < \binom{n}{pn} < G(p,n) \tag{18}$$

where

$$G(p,n) = \frac{1}{\sqrt{2\pi pqn}} p^{-pn} q^{-qn}.$$

¶19. Tail of the binomial distribution. The "tail" of the distribution $B(n,p)$ is the following sum

$$\sum_{i=\lambda n}^{n} \binom{n}{i} p^i q^{n-i}.$$

It is easy to see the following inequality:

$$\sum_{i=\lambda n}^{n} \binom{n}{i} p^i q^{n-i} \leq \binom{n}{\lambda n} p^{\lambda n}.$$

To see this, note that LHS is the probability of the event $A = \{$There are at least λn successsess in n coin tosses$\}$. For any choice x of λn out of n coin tosses, let B_x be the event that the chosen coin tosses are successes. Then RHS is the sum of the probability of B_x, over all x. Clearly $A = \cup_x B_x$. But the RHS may be an overcount because the events B_x need not be disjoint. We have the following upper bound [3]:

$$\sum_{i=\lambda n}^{n} \binom{n}{i} p^i q^{n-i} < \frac{\lambda q}{\lambda - p} \binom{n}{\lambda n} p^{\lambda n} q^{\mu n}$$

where $\lambda > p$ and $q = 1 - p$. This specializes to

$$\sum_{i=\lambda n}^{n} \binom{n}{i} < \frac{\lambda}{2\lambda - 1} \binom{n}{\lambda n}$$

where $\lambda > p = q = 1/2$.

¶20. Markov Inequality. Let X be a non-negative random variable. We have the trivial bound

$$\Pr\{X \geq 1\} \leq \mathtt{E}[X]. \tag{19}$$

For any real constant $c > 0$, $\Pr\{X \geq c\} = \Pr\{X/c \geq 1\} \leq \mathtt{E}[X/c] = \mathtt{E}[X]/c$. This proves[6] the so-called **Markov inequality**,

$$\Pr\{X \geq c\} \leq \frac{\mathtt{E}[X]}{c}. \tag{20}$$

Observe that the Markov inequality is trivial unless $\mathtt{E}[X]$ is finite and we choose $c > \mathtt{E}[X]$.

[6] Another proof uses the **Heaviside function** $H(x)$ that is the 0-1 function given by $H(x) = 1$ if and only if $x > 0$. We have the trivial inequality $H(X - c) \leq \frac{X}{c}$ Taking expections on both sides yields the Markov inequality since $\mathtt{E}[H(X - c)] = \Pr\{X \geq c\}$.

¶21. Chebyshev Inequality. It is also called the Chebyshev-Bienaymé inequality since it originally appeared in a paper of Bienaymé in 1853 [4, p. 73]. With any real $c > 0$,

$$\Pr\{|X| \geq c\} = \Pr\{X^2 \geq c^2\} \leq \frac{\mathtt{E}[X^2]}{c^2} \tag{21}$$

by an application of Markov inequality. Another form of this inequality (derived in exactly the same way) is

$$\Pr\{|X - \mathtt{E}[X]| \geq c\} = \Pr\{(X - \mathtt{E}[X])^2 \geq c^2\} \leq \frac{\mathtt{Var}(X)}{c^2}. \tag{22}$$

Sometimes $\Pr\{|X - \mathtt{E}[X]| \geq c\}$ is called the **tail probability** of X. By a trivial transformation of parameters, equation (22) can also written as

$$\Pr\{|X - \mathtt{E}[X]| \geq c\sqrt{\mathtt{Var}(X)}\} \leq \frac{1}{c^2}. \tag{23}$$

This form is useful in statistics because it bounds the probability of X deviating from its mean by some fraction of the standard deviation, $\sqrt{\mathtt{Var}(X)}$.

Let us give an application of Chebyshev's inequality:

Lemma 7. *Let X be a r.v. with mean $\mathtt{E}[X] = \mu \geq 0$.*
(a) Then

$$\Pr\{X = 0\} \leq \frac{\mathtt{Var}(X)}{\mu^2}.$$

(b) Suppose $X = \sum_{i=1}^{n} X_i$ where the X_i's are pairwise independent Bernoulli r.v.s with $\mathtt{E}[X_i] = p$ (and $q = 1 - p$) then

$$\Pr\{X = 0\} \leq \frac{q}{np}.$$

Proof. (a) Since $\{X = 0\} \subseteq \{|X - \mu| \geq \mu\}$, we have

$$\Pr\{X = 0\} \leq \Pr\{|X - \mu| \geq \mu\} \leq \frac{\mathtt{Var}X}{\mu^2}$$

by Chebyshev.
(b) It is easy to check that $\mathtt{Var}(X_i) = pq$. Since the X_i's are independent, we have $\mathtt{Var}(X) = npq$. Also $\mathtt{E}[X] = \mu = np$. Plugging into the formula in (a) yields the claimed bound on $\Pr\{X = 0\}$. **Q.E.D.**

Part (b) is useful in reducing the error probability in a certain class of randomized algorithms called RP-algorithms. The outcome of an RP-algorithm A may be regarded as a Bernoulli r.v. X_i which has value 1 or 0. If $X_i = 1$, then the algorithm has no error. If $X_i = 0$, then the probability of error is at most p $(0 \leq p < 1)$. We can reduce the error probability in RP-algorithms by repeating its computation n times and output 0 iff each of the n repeated computations output 0. Then part (b) bounds the error probability of the iterated computation. We will see several such algorithms later (e.g., primality testing in §XIX.2).

¶22. Jensen's Inequality. Let $f(x)$ be a real function. By definition, f is **convex** means that for all n,

$$f(\sum_{i=1}^{n} p_i x_i) \leq \sum_{i=1}^{n} p f(x_i)$$

where $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$. If X and $f(X)$ are random variables then

$$f(\mathtt{E}[X]) \leq \mathtt{E}[f(X)].$$

Let us prove this for the case when X has takes on finitely many values x_i with probability p_i. Then $\mathtt{E}[X] = \sum_i p_i x_i$ and

$$f(\mathtt{E}[X]) = f(\sum_i p_i x_i) \leq \sum_i p_i f(x_i) = \mathtt{E}[f(X)].$$

For instance, if $r \geq 1$ then $\mathtt{E}[|X|^r] \geq (\mathtt{E}[|X|])^r$.

Exercises

Exercise 7.1: Verify the equation (18). ♢

Exercise 7.2: Describe the class of non-negative random variables for which Markov's inequality is tight. ♢

Exercise 7.3: Chebyshev's inequality is the best possible. In particular, show an X such that $\Pr\{|X - \mathtt{E}[X]| > e\} = \mathtt{Var}(X)/e^2$. ♢

§8. Chernoff Bounds

Suppose we wish an upper bound on the probability $\Pr\{X \geq c\}$ where X is an arbitrary r.v.. To apply Markov's inequality, we need to convert X to a non-negative r.v. One way is to use the r.v. X^2, as in the proof of Chebyshev's inequality. The technique of Chernoff converts X to the Markov situation by using

$$\Pr\{X \geq c\} = \Pr\{e^X \geq e^c\}.$$

Since e^X is a non-negative r.v., we conclude from Markov's inequality (20) that

$$\Pr\{X \geq c\} \leq e^{-c}\mathtt{E}[e^X]. \tag{24}$$

We can further exploit this trick: for any positive number $t > 0$, we have $\Pr\{X \geq c\} = \Pr\{tX \geq tc\}$, and proceding as before, we obtain

$$\begin{aligned} \Pr\{X \geq c\} &\leq e^{-ct}\mathtt{E}[e^{tX}] \\ &= \mathtt{E}[e^{t(X-c)}]. \end{aligned}$$

Finally, the so-called Chernoff bound [1] is given by choosing t to minimize the right-hand side of this inequality. This proves:

Lemma 8 (Chernoff Bound). *For any r.v. X and real c,*

$$\Pr\{X \geq c\} \leq m(c). \tag{25}$$

where

$$m(c) = m_X(c) := \inf_{t>0} \mathtt{E}[e^{t(X-c)}]. \tag{26}$$

More generally, any bound that are derived from (25) is also called a Chernoff bound. We now derive some Chernoff bounds under various assumptions.

Let $X_1, \ldots, X_n$ be independent and

$$S = X_1 + \cdots + X_n.$$

It is easily verified that then $e^{tX_1}, \ldots, e^{tX_n}$ (for any constant t) are also independent. Then equation (15) implies

$$\mathtt{E}[e^{tS}] = \mathtt{E}[\prod_{i=1}^{n} e^{tX_i}] = \prod_{i=1}^{n} \mathtt{E}[e^{tX_i}].$$

(A) Suppose that, in addition, the $X_1, \ldots, X_n$ are i.i.d., and $m(c)$ is defined as in (26). This shows

$$\begin{aligned} \Pr\{S \geq nc\} &\leq e^{-nct}\mathtt{E}[e^{tS}], \quad (t > 0) \\ &\leq [m(c)]^n. \end{aligned}$$

This is a generalization of (25).

(B) Assume S has the distribution $B(n, p)$. It is not hard to compute that

$$m(c) = \left(\frac{p}{c}\right)^c \left(\frac{1-p}{1-c}\right)^{1-c}. \tag{27}$$

Then for any $0 < \varepsilon < 1$:

$$\Pr\{S \geq (1-\varepsilon)np\} \leq \left(\frac{1}{1-\varepsilon}\right)^{(1-\varepsilon)np} \left(\frac{1-p}{1-(1-\varepsilon)p}\right)^{n-(1+\varepsilon)np}.$$

We still need to make this bound more convenient for application:

$$\begin{aligned} \Pr\{S \geq (1+\varepsilon)np\} &\leq \exp(-\varepsilon^2 np/3) && (28) \\ \Pr\{S \leq (1-\varepsilon)np\} &\leq \exp(-\varepsilon^2 np/2) && (29) \\ & && (30) \end{aligned}$$

Need the $\leq$ and $\geq$ version of Chernoff bound... INCOMPLETE

(C) Now suppose the X_i's are independent Bernoulli variables where $\Pr\{X_i = 1\} = p_i$ $(0 \leq p_i \leq 1)$ and $\Pr\{X_i = 0\} = 1 - p_i$ for each i. Then

$$\mathtt{E}[X_i] = p_i, \qquad \mu := \mathtt{E}[S] = \sum_{i=1}^{n} p_i.$$

Fix any $\delta > 0$. Then

$$\begin{aligned} \Pr\{S \geq (1+\delta)\mu\} &\leq m((1+\delta)\mu) \\ &= \inf_{t>0} \mathtt{E}[e^{t(X-(1+\delta)\mu)}] \\ &= \inf_{t>0} \frac{\mathtt{E}[e^{tX}]}{e^{(1+\delta)\mu}}. \end{aligned}$$

¶23. Estimating a Probability and Hoeffding Bound. Consider the natural problem of estimating p $(0 < p < 1)$ where p is the probability that a given coin will show up heads in a toss. The obvious solution is to choose some reasonably large n, toss this coin n times, and estimate p by the ratio h/n where h is the number of times we see heads in the n coin tosses.

This problem is still not well-defined since we have no constraints on n. So assume our goal is to satisfy the bound

$$\Pr\{|p - (h/n)| > \delta\} \leq \varepsilon \tag{31}$$

where δ is the **precision parameter** and ε is a bound on the **error probability**. Given δ and ε, $(0 < \delta, \varepsilon < 1)$, we now have a well-defined problem. This problem seems to be solved by the Chernoff bound (B) in (28) where $S = X_1 + \cdots X_n$ is now interpreted to be h. Then

$$\{|p - (h/n)| > \delta\} = \{|np - h| > n\delta\} = \{|np - S| > n\delta\}$$

If we substitute δ with $p\varepsilon$, then we obtain

$$\begin{aligned} \Pr\{|p - (h/n)| > \delta\} &= \Pr\{|np - S| > n\delta\} \\ &= \Pr\{|S - np| > np\varepsilon\} \\ &\leq 2\exp(- \end{aligned}$$

The problem is that the p that we are estimating appears on the right hand side. Instead, we need the following Hoeffding bound:

$$\Pr\{S > np + \delta\} \leq \exp -n\delta^2/2 \tag{32}$$
$$\Pr\{S < np - \delta\} \leq \exp -n\delta^2/2 \tag{33}$$
$$\Pr\{|S < np| > \delta\} \leq 2\exp -n\delta^2/2 \tag{34}$$

Comparing the usual Chernoff bounds with the Hoeffding bound, we see that the former bound the relative error in the estimate while the latter concerns absolute error.

For a survey of Chernoff Bounds, see T. Hagerub and C. Rüb, "A guided tour of Chernoff Bounds", **Information Processing Letters** 33(1990)305–308.

Exercises

Exercise 8.1: Verify the equation (27). ♢

Exercise 8.2: Obtain an upper bound on $\Pr\{X \leq c\}$ by using Chernoff's technique. HINT: $\Pr\{X \leq c\} = \Pr\{tX \geq tc\}$ where $t < 0$. ♢

Exercise 8.3: Show the following:
i) Bonferroni's inequality,

$$\Pr(AB) \geq \Pr(A) + \Pr(B) - 1.$$

ii) Boole's inequality,

$$\Pr(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} \Pr(A_i).$$

(This is trivial, and usually used without acknowledgement.) iii) For all real x, $e^{-x} \geq 1 - x$ with equality only if $x = 0$.
iv) $1 + x < e^x < 1 + x + x^2$ which is valid for $|x| < 1$.

♢

Exercise 8.4: Kolmogorov's inequality: let $X_1, \ldots, X_n$ be mutually independent with expectation $\mathtt{E}[X_i] = m_i$ and variance $\mathtt{Var}(X_i) = v_i$. Let $S_i = X_1 + \cdots + X_i$, $M_i = \mathtt{E}[S_i] = m_1 + \cdots + m_i$ and $V_i = \mathtt{Var}(S_i) = v_1 + \cdots + v_i$. Then for any $t > 0$, the probability that the n inequalities

$$|S_i - M_i| < tV_n, \qquad i = 1, \ldots, n,$$

holds simultaneously is at least $1 - t^{-2}$. ♢

Exercise 8.5: We want to process a sequence of **requests** on a single (initially empty) list. Each request is either an insertion of a key or the lookup on a key. The probability that any request is an insertion is p, $0 < p < 1$. The cost of an insertion is 1 and the cost of a lookup is m if the current list has m keys. After an insertion, the current list contains one more key.
(a) Compute the expected cost to process a sequence of n requests.
(b) What is the *approximate* expected cost to process the n requests if we use a binary search tree instead? Assume that the cost of insertion, as well as of lookup, is $\log_2(1 + m)$ where m is the number of keys in the current tree. NOTE: If L is a random variable (say, representing the length of the current list), assume that $\mathtt{E}[\log_2 L] \approx \log_2 \mathtt{E}[L]$, (*i.e.*, the expected value of the log is approximately the log of the expected value).
(c) Let p be fixed, n varying. Describe a rule for choosing between the two datastructures. Assuming $n \gg 1 \gg p$, give some rough estimates (assume $\ln(n!)$ is approximately $n \ln n$ for instance).
(d) Justify the approximation $\mathtt{E}[\log_2 L] \approx \log_2 \mathtt{E}[L]$ as reasonable. ♢

§9. Generating Functions

In this section, we assume that our r.v.'s are discrete with range $\mathbb{N} = \{0, 1, 2, \ldots\}$.

This powerful tool of probabilistic analysis was introduced by Euler (1707-1783). If $a_0, a_1, \ldots,$ is a denumerable sequence of numbers, then its **(ordinary) generating function** is the power series

$$G(t) := a_0 + a_1 t + a_2 t^2 + \cdots = \sum_{i=0}^{\infty} a_i t^i.$$

If $a_i = \Pr\{X = i\}$ for $i \geq 0$, we also call $G(t) = G_X(t)$ the **generating function of** X. We will treat $G(t)$ purely formally, although under certain circumstances, we

can view it as defining a real (or complex) function of t. For instance, if $G(t)$ is a generating function of a r.v. X then $\sum_{i\geq 0} a_i = 1$ and the power series converges for all $|t| \leq 1$. The power of generating functions comes from the fact that we have a compact packaging of a potentially infinite series, facilitating otherwise messy manipulations. Differentiating (formally),

$$\begin{aligned} G'(t) &= \sum_{i=1}^{\infty} i a_i t^{i-1}, \\ G''(t) &= \sum_{i=2}^{\infty} i(i-1) a_i t^{i-2}. \end{aligned}$$

If $G(t)$ is the generating function of X, then

$$G'(1) = \mathtt{E}[X], \qquad G''(1) = \mathtt{E}[X^2] - \mathtt{E}[X].$$

It is easy to see that if $G_1(t) = \sum_{i\geq 0} a_i t^i$ and $G_2(t) = \sum_{i\geq 0} b_i t^i$ are the generating functions of independent r.v.'s X and Y then

$$G_1(t)G_2(t) = \sum_{i\geq 0} t^i \sum_{j=0}^{i} a_j b_{i-j} = \sum_{i\geq 0} t^i c_i$$

where $c_i = \Pr\{X+Y = i\}$. Thus we have: the product of the generating functions of two independent random variables X and Y is equal to the generating function of their sum $X+Y$. This can be generalized to any finite number of independent random variables. In particular, if $X_1, \ldots, X_n$ are n independent coin tosses (running example (E1)), then the generating function of X_i is $G_i(t) = q + pt$ where $q := 1 - p$. So the generating function of the r.v. $S_n := X_1 + X_2 + \cdots + X_n$ is

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

Thus, $\Pr\{S_n = i\} = \binom{n}{i} p^i q^{n-i}$ and S_n has the binomial distribution $B(n, p)$.

¶24. Moment generating function. The **moment generating function** of X is defined to be

$$\phi_X(t) := \mathtt{E}[e^{tX}] = \sum_{i\geq 0} a_i e^{it}.$$

This is sometimes more convenient then the ordinary generating function. Differentiating n times, we see $\phi_X^{(n)}(t) = \mathtt{E}[X^n e^{tX}]$ so $\phi^{(n)}(0)$ is the nth moment of X. For instance, if X is $B(n, p)$ distributed then $\phi_X(t) = (pe^t + q)^n$.

Exercises

Exercise 9.1:

(a) What is the generating function of the r.v. X where $\{X = i\}$ is the event that a pair of independent dice roll yields a sum of i $(i = 2, \ldots, 12)$?

(b) What is the generating function of $c_0, c_1, \ldots$ where $c_i = 1$ for all i? Where $c_i = i$ for all i? ♢

Exercise 9.2: Determine the generating functions of the following probability distributions: binomial, geometric, poisson. ♢

Exercise 9.3: Let $c_0 = 0$ and c_1 be some constant. For $n \geq 2$, consider the recurrence

$$c_n = \sum_{i=1}^{n} c_i c_{n-i}.$$

(a) If $G(X) = \sum_{i \geq 0} c_i X^i$ is the generating function of the c_n's, show that

$$G(X) = \frac{1 \pm \sqrt{1 - 4c_1 X}}{2}.$$

HINT: what is the connection between $G(X)^2$ and $G(X)$?
(b) Using the binomial theorem for $(1-x)^{1/2}$ determine the formula for c_n (as a function of c_1.
(c) What is the connection between c_i and the Catalan numbers (Lecture VI). ♢

Exercise 9.4: Compute the mean and variance of the binomial distributed, exponential distributed and Poisson distributed r.v.'s using generating functions. ♢

References

[1] H. Chernoff. A measure of asymptotic efficiency for tests of hypothesis based on sum of observations. *Ann. of Math. Stat.*, 23:493–507, 1952.

[2] K. L. Chung. *Elementary Probability Theory with Stochastic Processes*. Springer-Verlag, New York, 1979.

[3] W. Feller. *An introduction to Probability Theory and its Applications*. Wiley, New York, 2nd edition edition, 1957. (Volumes 1 and 2).

[4] D. E. Knuth. *The Art of Computer Programming: Fundamental Algorithms*, volume 1. Addison-Wesley, Boston, 2nd edition edition, 1975.

[5] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[6] A. N. Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Co., New York, 1956. Second English Edition.

[7] M. Li and P. M. B. Vitányi. *An introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, second edition, 1997.

[8] W. W. Peterson and J. E. J. Weldon. *Error-Correcting Codes*. MIT Press, 1975. 2nd Edition.