The most important questions of life are, for the most part, really only problems of probability. Strictly speaking one may even say that nearly all our knowledge is problematical; and in the small number of things which we are able to know with certainty, even in the mathematical sciences themselves, induction and analogy, the principal means for discovering truth, are based on probabilities, so that the entire system of human knowledge is connected with this theory.

Pierre-Simon Laplace. Introduction to Theorie Analytique des Probabilities. Oeuvres, t. 7. Paris, 1886, p. 5.

0.1 Syllabus

This syllabus is rough in its estimates of the time we'll spend on each topic. There are also some administrative details to be worked out, such as whether we have a TA and if not, how we'll get assignments graded.

Office hours: W3:00-4:00, Annenberg 317.

I'll be traveling a few times this term and will have to reschedule some classes. I'll try to use the Friday 10:30-12:00 time slot.

$\sim$ 1 week: Discrete probability: examples and basics. The probabilistic method and combinatorial applications. Randomized algorithms, derandomization by conditional expectations.

$\sim$ 2 weeks: "Fingerprinting", polynomial identity testing. Applications: file comparison, verifying matrix multiplication, verifying associativity. Bipartite matching. Hashing, AKS.

$\sim$ 3-4 weeks: Concentration of measure for fully or partially independent rvs. Fully and pairwise independent rvs. Chor-Goldreich error amplification. CLT, Large deviation bound (Chernoff / Bernstein), Chebychev bound. MIS and derandomization. Shannon coding theorem. Concentration of the number of prime factors of a random number. Set discrepancy. Unbalancing lights: upper and lower bounds. Expander graphs. Johnson-Lindenstrauss metric embedding. Bourgain metric embedding. k-wise independence and error-correcting codes. 4-wise independence and metric embedding? Talagrand inequality?

$\sim$ 2 weeks: Lovasz Local Lemma. Applications: Ramsey numbers, van der Waerden. Moser-Tardos algorithm.

$\sim$ 2 weeks: Randomized vs. Distributional Complexity. Game tree evaluation: upper and lower bounds. Karger's min-cut algorithm.

$\sim$ Unlikely to be reached: Linial-Saks graph partitioning. A. Kalai's sampling random factored numbers. Approximation of the permanent and self-reducibility. Equivalence of approximate counting and approximate sampling. ε-biased k-wise independent spaces. #DNF-approximation (Karp-Luby-Madras). Shamir secret sharing. An interactive proof for a problem only known to be in coNP: graph non-isomorphism. Markov Chain Monte Carlo. Weighted sampling.

1 Lecture 1, September 29, 2014

1.1 Appetizers

1. You have aerial photographs of a country, with known scale. Estimate the total length of all the roads.
2. N gentlemen check their hats in the lobby of the opera, but after the performance the hats are handed back at random. How many men, on average, get their own hat back?
3. The coins-on-dots problem: On the table before us are 10 dots, and in our pocket are 10 nickels. Prove the coins can be placed on the table (no two overlapping) in such a way that all the dots are covered.
4. Coupon collector.
5. Birthday Paradox. I just remind you of this: a class of just 23 students has even odds of some common birthday. (Supposing birthdates are uniform on 365 possibilities.) The exact calculation is

   $$\Pr(\text{some common birthday}) = 1 - \frac{365 \cdots 343}{365^{23}} \cong 0.507297$$

   but we'll provide a better way to understand this soon.
6. The envelope swap paradox: You're on a TV game show and the host offers you two identical-looking envelopes, each of which contains a check from the TV network, in your name. You pick whichever you like and take it, still unopened.

   Then the host explains: one of the checks is written for a sum of $\$N$ and the other is for $\$10N$. (and both for positive amounts). Now, he says, it's 50-50 whether you selected the small check or the big one. He'll give you a chance, if you like, to swap envelopes. It's a good idea for you to swap, he explains, because your expected net gain is (with $\$m$ representing the sum currently in hand):

   $$E(\text{gain}) = (1/2)(10m - m) + (1/2)(m/10 - m) = (81/20)m$$

   How can this be?
7. Consider a certain old-fashioned society in which parents prefer male offspring. Can a couple increase their expected fraction of sons by halting reproduction after the first?
8. $G(n, p)$ model. Threshold phenomenon for many properties such as connectivity, appearance of a Hamilton cycle, etc.
9. Unbalancing lights: there's an $n \times n$ grid of lightbulbs. For each bulb, at position (i, j), there is a switch b_{ij}; there is also a switch r_i on each row and a switch c_j on each column. The (i, j) bulb is lit if $b_{ij} + r_i + c_j$ is odd.

   What is the greatest $f(n)$ such that for any setting to the b_{ij}'s, you can set the row and column switches to light $n^2/2 + f(n)$ bulbs?

Notes. This course can be only an exposure to probability and its role in the theory of algorithms. We will stay focused on key ideas and examples; we will not be overconcerned with best bounds.

Problems will be assigned. There will be no exam. Lecture notes will be handed out after the fact. I'll be traveling twice for a week and we'll coordinate some make-up classes, perhaps on Friday at this same time.

Today is will be mostly about examples, basic probability and some combinatorics; it's not representative for the course, in which we'll spend a lot of time on algorithms. Despite the introductory nature of the lecture today, I assume this is not your first exposure to probability. Likewise I'll assume you have some basic familiarity with algorithms.

1.2 Some basics

What is a random variable?[1]

It is a *function* from a sample space into some range space.

A sample space, or probability space, is a triple $(\Omega, \tilde{\Omega}, \omega)$ where:

1. Ω is a set.
2. $\tilde{\Omega}$ is a "σ-field" on Ω, that is, a collection of subsets of Ω such that $\emptyset \in \tilde{\Omega}$, and $\tilde{\Omega}$ is closed under complement and countable intersection. It follows also that $\Omega \in \tilde{\Omega}$ and $\tilde{\Omega}$ is closed under countable union. The elements of $\tilde{\Omega}$ are called the *measurable sets*.
3. ω is a nonnegative-real-valued function on $\tilde{\Omega}$ such that
   (a) $\omega(\emptyset) = 0$
   (b) $\omega(\Omega) = 1$
   (c) For any $S \in \tilde{\Omega}$, $\omega(S) \geq 0$
   (d) For any $S, T \in \tilde{\Omega}$,

$$\omega(S) + \omega(T) = \omega(S \cap T) + \omega(S \cup T) \qquad \text{(modular identity)}$$

The range space can be any measure space W (that is, a set with an associated σ-field), and the random variable is required to be a measurable function, that is to say, pullbacks of measurable sets in W should be measurable sets in Ω. For instance, $W = \mathbb{R}$ (with the σ-field generated by intersections and complements of intervals), or $W =$ names of people participating in a lottery, or $W =$ deterministic algorithms for a certain computational problem).

For any (measurable) subset T of W, we associate the *event* $X \in T$, which we usually simply call the event T. This event has the probability $\Pr(X \in T)$ dictated by

$$\Pr(X \in T) = \omega(X^{-1}(T)).$$

It follows that these probabilities satisfy:

1. $\Pr(\emptyset) = 0$ ("something happens")
2. $\Pr(W) = 1$ ("only one thing happens")
3. $\Pr(T) \geq 0$
4. $\Pr(S) + \Pr(T) = \Pr(S \cap T) + \Pr(S \cup T)$

For the most part we will sidestep measure theory—one needs it to cure pathologies but we will be studying healthy patients. However I recommend Adams and Guillemin [2] or Billingsley [9].

It is hardly unusual to teach "probability without measure theory", but normally when people do this they suppress the role of the sample space in favor of abstract axioms of probability. For us the situation will be quite different. While starting out as a formality, soon explicit sample spaces will play a key role.

Indicator rvs
With every event T there is associated also a canonical "indicator random variable", sometimes written $[T]$, which equals 1 when $X \in T$ and equals 0 otherwise.

Joint distributions
There is a good reason we separated the *sample space* $(\Omega, \tilde{\Omega}, \omega)$ from the space W within which the random variable ranges. if we only cared about one rv, we wouldn't have to do this.

[1] For a philosophical and historical discussion of this question see Mumford in [33].

But consider for example the sample space in which Ω is the set of 36 ways in which two dice can roll, each outcome having probability $1/36$. On this sample space we can define various useful functions: e.g., $X_i =$ the value of die i ($i = 1, 2$); $Y = X_1 + X_2$.

Also, formally, given two random variables $X_1 : \Omega \to W_1$, $X_2 : \Omega \to W_2$ (where each W_i has associated with it a σ-field $\tilde{W}_i$), we can form the composite random variable $(X_1, X_2) : \Omega \to W_1 \times W_2$. Given a composite rv $(X_1, \ldots, X_n) : \Omega \to W_1 \times \ldots \times W_n$, its *marginals* are the n probability distributions on $W_1, \ldots, W_n$ defined by $\Pr(X_i \in A) = \Pr((X_1, \ldots, X_n) \in W_1 \times \ldots \times W_{i-1} \times A \times W_{i+1} \times \ldots \times W_n)$.

$X_1, \ldots$ are *independent* if for any finite $S = \{s_1, \ldots, s_n\}$ and all $A_{s_1} \in \tilde{W}_{s_1}, \ldots, A_{s_n} \in \tilde{W}_{s_n}$, $\Pr((X_{s_1}, \ldots, X_{s_n}) \in A_{s_1} \times \cdots \times A_{s_n}) = \Pr(X_{s_1} \in A_{s_1}) \cdots \Pr(X_{s_n} \in A_{s_n})$.

(Note that $\Pr((X_1, X_2) \in A_1 \times A_2)$ is just another way of writing $\Pr((X_1 \in A_1) \wedge (X_2 \in A_2))$.)

Consider the Dice Example. X_1 and X_2 are independent; X_1 and Y are not independent.

$X_1, \ldots : \Omega \to T$ are *independent and identically distributed (iid)* if they are independent and all marginals are identical. If T is finite and the marginals are the uniform distribution, we say that the rv's are uniform iid.

Shorthand
When the particular rv X is understood from context or immaterial, we abbreviate $\Pr(X \in A)$ by $\Pr(A)$.

The set $A - B$ is $A \cap (\neg B)$.

Conditional Probabilities are defined by

$$\Pr(X \in A | X \in B) = \frac{\Pr(X \in A \cap B)}{\Pr(X \in B)}$$

An old example. You meet Mr. Smith and find out that he has exactly two children, at least one of which is a girl. What is the probability that both are girls? Answer[2]: $1/3$.

Real-valued random variables; expectations
If X is a real-valued rv[3], its expectation (aka average, mean or first moment) is properly defined using the measure theory we are skipping, but can usually be expressed as follows:

$$E(X) = \lim_{h \to 0} \sum_{\text{integer} -\infty < j < \infty} jh \Pr(jh \leq X < (j+1)h) \tag{1}$$

The one caveat is that we require absolute convergence in order for the expectation to be well defined. Absolute convergence means that $\sup_{h>0} \sum_{-\infty<j<\infty} |jh| \Pr(jh \leq X < (j+1)h)$ is finite.

For $T \subseteq W$, considering its indicator rv, we have:

$$\Pr(X \in T) = \omega(X^{-1}(T)) = E([T]).$$

Now, if we have two real-valued rvs X, Y, we can form their sum rv $X + Y$. No matter the joint distribution of X and Y, we have:

$$E(X + Y) = E(X) + E(Y) \qquad \text{linearity of expectation}$$

for the simple reason that expectation is a first moment.

We have also countable additivity: let $X_1, \ldots$ be real-valued with expectations $E(X_i)$ and such that $\sum |E(X_i)| < \infty$. Then

$$E(\sum X_i) = \sum E(X_i).$$

[2] As usual in such examples we suppose that the sexes of the children are uniform iid.

[3] In this course, the real line $\mathbb{R}$ will always be considered to be equipped with the σ-field consisting of rays (a, ∞), rays $[a, \infty)$, and any set formed out of these by closing under the operations of complement and countable union. (In CS language, you use a finite-depth formula each node of which is one of these two operations, and whose leaves are the aforementioned rays.) It is often convenient to think of the "extended real line," which is the real line with ∞ and $-\infty$ adjoined.

Linearity of expectation already resolves several of our appetizers: 1, 2, 3, 4.

1: Enclose the photo in a disk of radius R and pass a line ℓ at a uniformly chosen angle, with displacement uniform between $\pm R$. A short stretch of road can be considered straight and if of length x, then when ℓ is at angle α to the road, there is probability $\frac{x}{2R}\sin\alpha$ of the two lines crossing. Since the orientation is uniform,

$$\Pr(\ell \text{ crosses the short stretch}) = \frac{1}{\pi}\int \frac{x}{2R}\sin\alpha \; d\alpha = \frac{x}{\pi R}$$

Now since ℓ crosses the short stretch at most once,

$$\Pr(\ell \text{ crosses the short stretch}) = E(\#\text{ times } \ell \text{ crosses the short stretch})$$

and so by linearity of expectation

$$\text{total roadlength} = \pi \cdot R \cdot E(\#\text{ times } \ell \text{ crosses roads})$$

2: Let $X_i = [\text{Gentleman } i \text{ receives his hat back}]$. These are certainly not independent, but no matter; $E(X_i) = 1/n$ and so $E(\sum X_i) = 1$.

3: I'll leave this as a puzzle.

4: Think of the coupons being sampled at times $1, 2, \ldots$. Let $Y_i =$ the first time at which we have seen i different kinds of coupons. Let $X_i = Y_i - Y_{i-1}$. After time Y_{i-1}, in each round there is probability $(n-i+1)/n$ that we see a new kind of coupon, until that finally happens. That is to say, X_i is geometrically distributed with parameter $p_i = (n-i+1)/n$:

$$\Pr(X_i = 1) = p_i, \quad \Pr(X_i = 2) = (1-p_i)p_i, \ldots \quad \Pr(X_i = m) = (1-p_i)^{m-1}p_i$$

$$E(X_i) = \sum_1^\infty m(1-p_i)^{m-1}p_i = \frac{1}{p_i}$$

$$E(Y_n) = \sum E(X_i) = \sum_1^n \frac{1}{p_i} = \sum_1^n \frac{n}{n-i+1} = n\sum_1^n \frac{1}{i} = nH_n = n(\log n + O(1))$$