Lecture 0 - Mathematical Background

This is a brief review of some mathematical tools, and especially probability theory, that we will use in this course.

At Harvard, much of this material (and more) is taught in Stat 110 "Introduction to Probability", CS20 "Discrete Mathemtics", and AM107 "Graph Theory and Combinatorics". Some good sources for this material are the lecture notes by Papadimitriou and Vazirani (see home page of Umesh Vaziarani), Lehman, Leighton and Meyer from MIT Course 6.042 "Mathematics For Computer Science" (Chapters 1-2 and 14 to 19 are particularly relevant). The mathematical tool we use most often is discrete probability. The "Probabilistic Method" book by Alon and Spencer is a great resource in this area. Also, the books of Mitzenmacher and Upfal and Prabhakar and Raghavan cover probability from a more algorithmic perspective. For an excellent popular discussion of some of the mathematical concepts we'll talk about, I can't recommend highly enough the book *"How Not to Be Wrong"* by Jordan Ellenberg.

Although knowledge of algorithms is not strictly necessary, it would be quite useful. Students who did not take CS124/CS125 might want to look at the books (1) Corman, Leiserson, Rivest and Smith, (2) Dasgupte, Papadimitriou and Vaziarni, or (3) Kleinberg and Tardos. We do not require prior knowledge of complexity or computability but some basic familiarity could be useful. Students who did not take CS121/CS125 might want to look at either Sipser's book (Intro to Theory of Computation) or the first 2 chapters of my book with Arora.

Mathematical Proofs

Arguably *the* mathematical prerequisite needed for this course is a certain level of comfort with mathematical proofs. Many students tend to think of mathematical proofs as a very formal object, like the proofs studied in school in geometry, consisting of a sequence of axioms and statements derived from them by very specific rules. In fact,

> *a proof is a piece of writing meant to convince human readers that a particular statement is true.*

(In this class, the particular humans you are trying to convince are me and the teaching fellows.)

To write a proof of some statement X you need to follow three steps:

1. Make sure that you completely understand the statement X.

2. Think about X until you are able to convince *yourself* that X is true.

3. Think how to present the argument in the clearest possible way so you can convince the reader as well.

Like any good piece of writing, a proof should be concise and not be overly formal or cumbersome. In fact, overuse of formalism can often be *detrimental* to the argument since it can mask weaknesses in the argument from both the writer and the reader. Sometimes students try to "throw the kitchen sink" at an answer trying to list all possibly relevant facts in the hope of getting partial credit. But a proof is a piece of writing, and a badly written proof will not get credit even if it contains some correct elements. It is better to write a clear proof of a partial statement. In particular, if you haven't been able to convince yourself that the statement is true, you should be honest about it and explain which parts of the statement you have been able to verify and which parts you haven't.

Example: The existence of infinitely many primes.

In the spirit of "do what I say and not what I do", I will now demonstrate the importance of conciseness by belaboring the point and spending several paragraphs on a simple proof, written by Euclid around 300 BC. Recall that a *prime number* is an integer $p > 1$ whose only divisors are p and 1. Euclid's Theorem is the following:

Theorem: There exist infinitely many primes.

Instead of simply writing down the proof, let us try to understand how we might figure this proof out. (If you haven't seen this proof before, or you don't remember it, you might want to stop reading at this point and try to come up with it on your own before continuing.) The first (and often most important) step is to understand what the statement means. Saying that the number of primes is infinite means that it is not finite. More precisely, this means that for every natural number k, there are not than k primes.

Now that we understand what we need to prove, let us try to convince ourselves of this fact. At first, it might seem obvious— since there are infinitely many natural numbers, and every one of them can be factored into primes, there must be infinitely many primes, right?

Wrong. Since we can compose a prime many times with itself, a finite number of primes can generate infinitely many numbers. Indeed the single prime 3 generates the infinite set of all numbers of the form 3^n. So, what we really need to show is that for every finite set of primes $\{p_1, \ldots, p_k\}$, there exists a number n that has a prime factor outside this set.

Now we need to start playing around. Suppose that we had just two primes p and q. How would we find a number n that is not generated by p and q? If you

try to draw things on the number line, you would see that there is always some *gap* between multiples of p and q in the sense that they are never consecutive. It is possible to prove that (in fact, it's not a bad exercise) but this observation already suggests a guess for what would be a number that is divisible by neither p nor q, namely $pq + 1$. Indeed, the remainder of $n = pq + 1$ when dividing by either p or q would be 1 (which in particular is not zero). This observation generalizes and we can set $n = pqr + 1$ to be a number that is divisable neither p, q nor r, and more generally $n = p_1 \cdots , p_k + 1$ is not divisable by $p_1, \ldots, p_k$.

Now we have convinced ourselves of the statement and it is time to think of how to write this down in the clearest way. One issue that arises is that we want to prove things truly from the definition of primes and first principles, and so not assume properties of division and remainders or even the existence of a prime factorization, without proving it. Here is what a proof could look like. We will prove the following two lemmas:

Lemma 1: For every integer $n > 1$, there exists a prime $p > 1$ that divides n.

Lemma 2: For every set of integers $p_1, \ldots, p_k > 1$, there exists a number n such that none of $p_1, \ldots, p_k$ divide n.

From these two lemmas it follows that there exist infinitely many primes, since otherwise if we let $p_1, \ldots, p_k$ be the set of all primes, then we would get a condtradiction as by combining Lemma 1 and Lemma 2 we would get a number n with a prime factor outside this set. We now prove the lemmas:

Proof of Lemma 1: Let $n > 1$ be a number, and let p be the smallest divisor of n that is larger than 1 (there exists such a number p since n divides itself). We claim that p is a prime. Indeed suppose otherwise there was some $1 < q < p$ that divides p. Then since $n = pc$ for some integer c and $p = qc'$ for some integer c' we'll get that $n = qcc'$ and hence q divides n in contradiction to the choice of p as the smallest divisor of n. QED

Proof of Lemma 2: Let $n = p_1 \cdots p_k + 1$ and suppose for the sake of contradiction that there exists some i such that $n = p_i \cdot c$ for some integer c. Then if we divide the equation $n - p_1 \cdots p_k = 1$ by p_i then we get c minus an integer on the lefthand side, and the fraction $1/p_i$ on the righthand side. QED

Some basic notation and concepts.

I will assume familiarity with basic notions of sets and operations on sets such as union (denoted $\cup$), intersection (denoted $\cap$), and set substraction (denoted $\setminus$). We denote by $|A|$ the size of the set A. I also assume familiarity with functions, and notions such as one-to-one (injective) functions and onto (surjective) functions. If f is a function from a set A to a set B, we denote this by $f : A \to B$. If f is one-to-one then this implies that $|A| \leq |B|$. If f is onto then $|A| \geq |B|$. If f is a permutation/bijection (e.g., one-to-one *and* onto) then this implies that $|A| = |B|$.

I also assume familiarity with *big-Oh notation*: If f, g are two functions from $\mathbb{N}$ to $\mathbb{N}$, then (1) $f = O(g)$ if there exists a constant c such that $f(n) \leq c \cdot g(n)$ for every sufficiently large n, (2) $f = \Omega(g)$ if $g = O(f)$, (3) $f = \Theta(g)$ is $f = O(g)$ and $g = O(f)$, (4) $f = o(g)$ if for every $\epsilon > 0$, $f(n) \leq \epsilon \cdot g(n)$ for every sufficiently large n, and (5) $f = \omega(g)$ if $g = o(f)$.

To emphasize the input parameter, I often write $f(n) = O(g(n))$ instead of $f = O(g)$, and use similar notation for $o, \Omega, \omega, \Theta$. While this is only an imprecise heuristic, when you see a statement of the form $f(n) = O(g(n))$ you can often replace it in your mind by the statement $f(n) \leq 1000g(n)$ while the statement $f(n) = \Omega(g(n))$ can often be thought of as $f(n) \geq 0.001g(n)$.

If n is an integer, then we denote by $a \pmod n$ the remainder of a when divided by n. $a \pmod n$ is the number $r \in \{0, \ldots, n-1\}$ such that $a = kn + r$ for some integer k. It will be very useful that $a \pmod n + b \pmod n = (a + b) \pmod n$ and $a \pmod n \cdot b \pmod n = (a \cdot b) \pmod n$ and so modular arithmetic inherits all of the rules (associativity, commutativity etc..) of integer arithmetic. If a, b are positive integers then $gcd(a, b)$ is the largest integer that divides both a and b. It is known that for every a, b there exist (not necessarily positive) integers x, y such that $ax + by = gcd(a, b)$ (it's a good exercise to prove this on your own). In particular, if $gcd(a, n) = 1$ then there exists a *modular inverse* for a which is a number b such that $ab = 1 \pmod n$. We sometimes write b as $a^{-1} \pmod n$.

For a set S, we denote by S^k the set of k-tuples from S. The most common case would be for $S = \{0, 1\}$ in which case S^k is the set of 0/1-valued strings of length k. The set S^* is defined as the set of all finite tuples from S. (i.e., $S^* = S^1 \cup S^2 \cup \cdots$). If $a, b \in \{0, 1\}$, then the AND,OR,XOR and NOT operations are defined respectively by $a \wedge b = ab$, $a \vee b = a + b - ab$, $a \oplus b = a + b \pmod 2$, and $\neg a = 1 - a$. Every function $f : \{0, 1\}^n \to \{0, 1\}$ can be obtained by combining these operations.

In later parts of the course we will need the notions of matrices, vectors, matrix multiplication and inverse, determinant, eigenvalues, and eigenvectors. These can be picked up in any basic text on linear algebra. In some parts we might also use some basic facts of group theory (finite groups only, and mostly only commutative ones). These also can be picked up as we go along, and a prior course on group theory is not necessary.

Probability and Sample spaces

Perhaps the main mathematical background needed in cryptography is probability theory since, as we will see, there is no secrecy without randomness. Luckily, we only need fairly basic notions of probability theory and in particular only probability over finite sample spaces. If you have a good understanding of what

happens when we toss k random coins, then you know most of the probability you'll need.

For every probabilistic experiment (for example, tossing a coin or throwing 3 dice) the set of all possible results of the experiment is called a *sample space*. For example, if the experiment is to toss a single coin and see if the result is "heads" or "tails" then the sample space is the set $\{H, T\}$, or equivalently (if we denote heads by 1 and tails by 0) the set $\{0, 1\}$. As another example, consider the experiment of tossing three coins. In this case there are 8 possible results and hence the sample space is $\{000, 001, 010, 011, 100, 101, 110, 111\}$. Each element in the sample space gets chosen with probability $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2^3} = \frac{1}{8}$.

Our sample space S will always be finite, and often it will be the case that all elements s in S are equally likely. In fact, in many cases our sample space S will be the 2^k-sized set $\{0,1\}^k$ corresponding to tossing k coins, with the probability of any particular k-length sequence s of heads and tails (or zeroes and ones) being 2^{-k}. We will use the notation $x \leftarrow_R \{0,1\}^k$ to denote that x is sampled uniformly at random from this sample space, and sometimes also use U_k to denote the same distribution.

Events

An *event* is a subset of the sample space. The probability that an event happens is the probability that the result of the experiment will fall inside that subset. For example, if we consider the sample space of tossing 101 coins, then we can denote by A the event that most of the coins came up tails — at most 50 of the coins come up "heads". In other words, A is the set of all length-101 strings with at most 50 ones. We denote the probability that an event A occurs by $\Pr[A]$. For example, in this case we can write

$$\Pr_{x \leftarrow_R \{0,1\}^{101}}[\#\text{ of 1's in } x \leq 50] = \frac{1}{2}$$

Proof: Let $A = \{x : \#\text{ of 1's in } x \leq 50\}$ as above. Let f be the function that flips all the bits of x from 1 to 0 and vice versa. Then f is a one-to-one and onto function from A to its complement $\overline{A} = \{0,1\}^{101} \setminus A$, meaning that $|A| = |\overline{A}|$ and since the A and $\overline{A}$ are disjoint sets whose union is

$$\Pr[A \cup A'] \leq \Pr[A] + \Pr[A']$$

We omit the (very simple) proof– can you see why this is true?

$\{0,1\}^{101}$ it follows that $|A| = \frac{2^{101}}{2}$. QED

Union Bound

If A and A' are events over the same sample space then another way to look at the probability that *either* A or A' occurs is to say that this is the probability that the event $A \cup A'$ (the union of A and A') occurs. A very simple but useful bound is that this probability is *at most* the sum of the probability of A and the probability of A'. This is called the *union bound*

Theorem (Union bound): If S is a sample space and $A, A' \subseteq S$ are two events over S. Then,

Note that that there are examples of A and A' such that $\Pr[A \cup A']$ is strictly less than $\Pr[A] + \Pr[A']$. For example, this can be the case if A and A' are the same set (and hence $A \cup A' = A$). If A and A' are *disjoint* (i.e., mutually exclusive) then $\Pr[A \cup A'] = \Pr[A] + \Pr[A']$.

Random Variables

A random variable is a function that maps elements of the sample space to another set (often, but not always, to the set $\mathbb{R}$ of real numbers). For example, in the case of the uniform distribution over $\{0,1\}^{101}$, we can define the random variable N to denote the number of ones in the string chosen. That is, for every $x \in \{0,1\}^{101}$, $N(x)$ is equal to the number of ones in x. Thus, the event A we considered before can be phrased as the event that $N \leq 50$ and the formula above can be phrased as

$$\Pr_{x \leftarrow_R \{0,1\}^{101}}[N(x) \leq 50] = \frac{1}{2}$$

For the remainder of this lecture, we will only consider **real** random variables (that is random variables whose output is a **real number**).

Expectation

The *expectation* of a random variable is its weighted average. That is, it is the average value it takes, when the average is weighted according to the probability measure on the sample space. Formally, if N is a random variable on a sample space S (where for every $x \in S$, the probability that x is obtained is given by p_x) then the expectation of N, denoted by $\mathbb{E}[N]$ is defined as follows:

$$\mathbb{E}[N] := \sum_{x \in S} N(x) \cdot p_x$$

For example, if the experiment was to choose a random U.S. citizen (and hence the sample space is the set of all U.S. citizens) and we defined the random

variable H to be the height of the person chosen, then the expectation of H (denoted by $\mathbb{E}[H]$) is simply the average height of a U.S. citizen.

There can be two different random variables with the same expectation. For example, consider the sample space $\{0,1\}^{101}$ with the uniform distribution, and the following two random variables:

- N is the random variable defined above: $N(x)$ is the number of ones in x.
- M is defined as follows: if x is the all ones string (that is $x = 1^{101}$) then $M(x) = 50.5 \cdot 2^{101}$. Otherwise (if $x \neq 1^{101}$) then $M(x) = 0$.

The expectation of N equals 50.5 (this follows from the linearity of expectation, see below).

The expectation of M is also 50.5: with probability 2^{-101} it will be $2^{101} \cdot 50.5$ and with probability $1 - 2^{-101}$ it will be 0.

Note that even though the average of M is 50.5, the probability that for a random x, $M(x)$ will be close to 50.5 or even bigger than zero is very very small. This is similar to the fact that if Bill Gates is in a room with any group of 99 people, no matter how pool, then the average worth of a random person in this room is more than \$100M even though with probability 0.99 a random person in the room will be worth much less than that amount. Hence the name "expectation" is somewhat misleading.

In contrast, we will see from the Chernoff bound below, that for a random string x, even though it will never have $N(x)$ equal to exactly 50.5 (after all, $N(x)$ is always a whole number), with high probability $N(x)$ will be close to 50.5.

The fact that two very different variables can have the same expectation means that if we know the expectation it does not give us *all* the information about the random variable but only *partial* information.

Linearity of expectation. The expectation function has a very useful property which is that it is a *linear function*. That is, if N and M are random variables over the same sample space S, then we can define the random variable $N + M$ in the natural way: for every $x \in S$, $(N + M)(x) = N(x) + M(x)$. It turns out that $\mathbb{E}[N + M] = \mathbb{E}[N] + \mathbb{E}[M]$. For every fixed number c and random variable N we define the random variable cN in the natural way: $(cN)(x) = c \cdot N(x)$ It turns out that $\mathbb{E}[cN] = c\mathbb{E}[N]$.

For example, the random variable N above is equal to $X_1 + \cdots + X_{101}$ with X_i equalling the i^{th} bit of the chosen string. Since $\mathbb{E}[X_i] = (1/2) \cdot 0 + (1/2) \cdot 1 = 1/2$, $\mathbb{E}[N] = 101 \cdot (1/2) = 50.5$.

Deviation bounds

As we saw above, sometimes we want to know not just the expectation of a random variable but also the probability that the variable is close to (or at

least not too far from) its expectation. Bounds on this probability are often called "tail bounds" or "deviation bounds". We now discuss the most common such bounds. Their general flavor is that the more you know about the random variable (which usually amounts to knowing *higher moments* of the random variables - expectations of powers larger than one) the better you can bound its deviation from the expectation.

Markov Inequality

The simplest tail bound is *Markov's inequality*, which is a one-sided inequality. It says that with high probability a non-negative random variable is never much larger than its expectation. (Note that the random variable M defined above was an example of a non-negative random variable that with high probability is much *smaller* than its expectation.) That is, it is the following theorem:

Theorem (Markov's Inequality): Let X be a random variable over a sample space S such that for all $s \in S$, $X(s) \geq 0$. Let $k \geq 1$. Then,

$$\Pr[X \geq k\mathbb{E}[X]] \leq \frac{1}{k}$$

proof: Denote $\mu = \mathbb{E}[X]$ and let $A = \{s \in S \mid X(s) \geq k\mu\}$. By the definition of expectation

$$\mathbb{E}[X] = \sum_{s \in S} X(s) \Pr[s] = \sum_{s \in A} X(s) \Pr[s] + \sum_{s \notin A} X(s) \Pr[s] \ .$$

Since the second term is non-negative

$$\mu \geq \sum_{s \in A} X(s) \Pr[s] \ .$$

However, we know that for each $s \in A$, $X(s) \geq k\mu$ and hence

$$\sum_{x \in A} X(s) \Pr[s] \geq k\mu \sum_{s \in A} \Pr[s] = k\mu \Pr[A]$$

. Combining these two equations we get $\mu \geq k\mu \Pr[A]$ or $\Pr[A] \leq 1/k$ which is what we wanted to prove. QED

Variance and Chebychev inequality

We already noted that the distance from the expectation is an interesting parameter. Thus, for a random variable X with expectation μ we can define a new random variable $\tilde{X}$ which to be the distance of X from its expectation. That is, for every $s \in S$, we define $\tilde{X}(s) = |X(s) - \mu|$. (Recall that $|\cdot|$ denotes the absolute value.) It turns out that it is hard to work with $\tilde{X}$ and so we look at the variable $\tilde{X}^2$, which is equal to $(X - \mu)^2$. We define the *variance* of a random variable X to be equal to the expectation of $\tilde{X}^2$. That is, for X with $\mathbb{E}[X] = \mu$,

$$Var[X] := \mathbb{E}[\tilde{X}^2] = \mathbb{E}[(X - \mu)^2]$$

In other words $Var[X]$ is defined to be $\mathbb{E}[(X - \mathbb{E}[X])^2]$.

We define the *standard deviation* of X to be the square root of $Var[X]$.

If we have a bound on the variance then we can have a better tail bound on the variables:

Theorem (Chebyshev's inequality): Let X be a random variable over S with expectation μ and standard deviation σ. Let $k \geq 1$. Then,

$$\Pr[|X - \mu| \geq k\sigma] \leq 1/k^2$$

Proof: The variable $Y = (X - \mu)^2$ is non-negative and has expectation $Var(X) = \sigma^2$. Therefore, by Markov inequality,

$$\Pr[(X - \mu)^2 \geq k^2\sigma^2] = \Pr[Y \geq k^2\sigma^2] \leq 1/k^2 \ .$$

However, since $|X - \mu| \geq k\sigma$ holds if and only if $(X - \mu)^2 \geq k^2\sigma^2$ the probability of these two events is identical. Thus $\Pr[|X - \mu| \geq k\sigma] \leq 1/k^2$. QED

Conditional probabilities and independence

Let A be some event over a sample space S (with $\Pr[A] > 0$). By a probability *conditioned on* A we mean the probability of some event, assuming that we already know that A happened. For example if S is our usual sample space of uniform choices over $\{0,1\}^{101}$ and A is the event that the first coin turned out head, then the conditional space is the space of all length-101 strings whose first bit is 1.

Formally this is defined in the natural way: we consider A as a sample space by inheriting the probabilities from S (and normalizing so the probabilities will sum up to one). That is, for every $x \in A$ we define $\Pr[x|A]$ (the probability that x is

chosen conditioned on A) to be $\Pr[x]/\Pr[A]$. For an event B we define $\Pr[B|A]$ (the probability that B happens conditioned on A) to be $\sum_{x \in A \cap B} \Pr[x|A] = \Pr[A \cap B]/\Pr[A]$.

Independent events. We say that B is *independent* from A if $\Pr[B|A] = \Pr[B]$. That is, knowing that A happened does not give us any new information on the probability that B will happen. By plugging the formula for $\Pr[B|A]$ we see that B is independent from A if and only if

$$\Pr[B \cap A] = \Pr[A] \Pr[B]$$

This means that B is independent from A iff A is independent from B and hence we simply say that A and B are independent events.

For example, if, as above, A is the event that the first coin toss is heads and B is the event that the second coin toss is heads then these are independent events. In contrast if C is the event that the number of heads is at most 50 then C and A are *not* independent (since knowing that A happened increases somewhat the chances for C).

If we have more than two events then it's a bit more messy: we say that the events $A_1, \ldots, A_n$ are *mutually independent* if not only $\Pr[A_1 \cap A_2 \cap \cdots \cap A_n] = \Pr[A_1] \cdots \Pr[A_n]$ but also this holds for every subset of $A_1, \ldots, A_n$. That is, for every subset I of the numbers $\{1, \ldots, n\}$,

$$\Pr[\cap_{i \in I} A_i] = \prod_{i \in I} \Pr[A_i]$$

Independent random variables. We say that U and V are **independent random variables** if for every possible values u and v, the events $U = u$ and $V = v$ are independent events or in other words $\Pr[U = u \text{ and } V = v] = \Pr[U = u] \Pr[V = v]$. We say that $U_1, \ldots, U_n$ are a collection of independent random variables if for all values $u_1, \ldots, u_n$, the events $U_1 = u_1, \ldots, U_n = u_n$ are mutually independent.

The Chernoff Bound

Suppose that 60% of a country's citizens prefer the color blue over red. A poll is the process of choosing a random citizen and finding his or her favorite color. Suppose that we do this n times and we define the random variable X_i to be 0 if the color of the i^{th} person chosen is red and 1 if it is blue. Then, for each i the expectation $\mathbb{E}[X_i]$ is 0.6, and by linearity of expectation $\mathbb{E}[\sum_{i=1}^{n} X_i] = 0.6n$. The estimate we get out of this poll for the fraction of blue-preferrers is $\frac{\sum X_i}{n}$ and we would like to know how close this is to the real fraction of the population (i.e.,

0.6). In other words, for any $\epsilon > 0$, we would like to know what is the probability that our estimate will be ϵ off from the real value, i.e., that $|\frac{\sum X_i}{n} - 0.6| > \epsilon$.

It turns out that in this case we have a very good bound on the deviation of $\sum X_i$ from its expectation, and this is because all of the X_i's are independent random variables (since in each experiment we draw a new random person independently of the results of previous experiments). This is the Chernoff bound, which we state here in a simplified form:

Theorem (Chernoff bound): Let $X_1, \ldots, X_n$ be independent random variables with $0 \leq X_i \leq 1$ and $\mathbb{E}[X_i = \mu]$. Then,

$$\Pr\left[\left|\frac{\sum X_i}{n} - \mu\right| > \epsilon\right] < 2^{-\epsilon^2 n/4}$$

We omit the proof that can be found in many of the texts mentioned above, though see the exercises for a proof of an important special case.

Exercises

The following exercises will be part of the first problem set in the course, so you can get a head start by working on them now.

1. In the following exercise X, Y denote random variables over some sample space S. You can assume that the probability on S is the uniform distribution— every point s is output with probability $1/|S|$. Thus $\mathbb{E}[X] = (1/|S|)\sum_{s \in S} X(s)$. We define the variance and standard deviation of X and Y as above (e.g., $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$ and the standard deviation is the square root of the variance).
    1. Prove that $Var[X]$ is always non-negative.
    2. Prove that $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
    3. Prove that always $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$.
    4. Give an example for a random variable X such that $\mathbb{E}[X^2] \neq \mathbb{E}[X]^2$.
    5. Give an example for a random variable X such that its standard deviation is *not equal* to $\mathbb{E}[|X - \mathbb{E}[X]|]$.
    6. Give an example for two random variables X, Y such that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
    7. Give an example for two random variables X, Y such that $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$.
    8. Prove that if X and Y are independent random variables (i.e., for every x, y, $\Pr[X = x \wedge Y = y] = \Pr[X = x]\Pr[Y = y]$) then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ and $Var[X + Y] = Var[X] + Var[Y]$.

2. Suppose that H is chosen to be a random function mapping the numbers $\{1,\ldots,n\}$ to the numbers $\{1,..,m\}$. That is, for every $i \in \{1,\ldots,n\}$, $H(i)$ is chosen to be a random number in $\{1,\ldots,m\}$ and that choice is done independently for every i. For every $i < j \in \{1,\ldots,n\}$, define the random variable $X_{i,j}$ to equal 1 if there was a *collision* between $H(i)$ and $H(j)$ in the sense that $H(i) = H(j)$ and to equal 0 otherwise.

    1. For every $i < j$, compute $\mathbb{E}[X_{i,j}]$.
    2. Define $Y = \sum_{i<j} X_{i,j}$ to be the total number of collisions. Compute $\mathbb{E}[Y]$ as a function of n and m. In particular your answer should imply that if $m < n^2/1000$ then $\mathbb{E}[Y] > 1$ and hence in expectation there should be at least one collision and so the function H will not be one to one.
    3. Prove that if $m > 1000 \cdot n^2$ then the probability that H is one to one is at least 0.9.
    4. Give an example of a random variable Z (unrelated to the function H) that is always equal to a non-negative integer, and such that $\mathbb{E}[Z] \geq 1000$ but $\Pr[Z > 0] < 0.001$.
    5. Prove that if $m < n^2/1000$ then the probability that H is one to one is at most 0.1.

3. In this exercise we we will work out an important special case of the Chernoff bound. You can take as a given the following facts:

    1. The number of $x \in \{0,1\}^n$ such that $\sum x_i = k$ is $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.
    2. Stirling's approximation formula: for every $n \geq 1$,

    $$\sqrt{2\pi n}\left(\tfrac{n}{e}\right)^n \leq n! \leq 2\sqrt{2\pi n}\left(\tfrac{n}{e}\right)^n$$

    where $e = 2.7182\ldots$ is the base of the natural logarithm.

    Do the following:

    1. Prove that for every n, $\Pr_{x \leftarrow_R \{0,1\}^n}[\sum x_i \geq 0.6n] < 2^{-n/1000}$

    The above shows that if you were given a coin of bias at least 0.6, you should only need some constant number of samples to be able to reject the "null hypothesis" that the coin is completely unbiased with extremely high confidence. In the following somewhat more challenging questions (which can be considered as bonus exercise) we try to show a converse to this:

    1. Let P be the uniform distribution over $\{0,1\}^n$ and Q be the $1/2+\epsilon$-biased distribution corresponding to tossing n coins in which each one has a probability of $1/2+\epsilon$ of equalling 1 and probability $1/2-\epsilon$ of equalling 0. Namely the probability of $x \in \{0,1\}^n$ according to Q is equal to $\prod_{i=1}^n (1/2 - \epsilon + 2\epsilon x_i)$.

1. Prove that for every threshold θ between 0 and n, if $n < 1/(100\epsilon)^2$ then the probabilities that $\sum x_i \leq \theta$ under P and Q respectively differ by at most 0.1. Therefore, one cannot use the test whether the number of heads is above or below some threshold to reliably distinguish between these two possibilities unless the number of samples n of the coins is at least some constant times $1/\epsilon^2$.

2. Prove that for *every* function F mapping $\{0,1\}^n$ to $\{0,1\}$, if $n < 1/(100\epsilon)^2$ then the probabilities that $F(x) = 1$ under P and Q respectively differ by at most 0.1. Therefore, if the number of samples is smaller than a constant times $1/\epsilon^2$ then there is simply *no test* that can reliably distinguish between these two possiblities.