1 Random variables

We defined a probability space in terms of a probability assigned to each element in a set. Events were just subsets of this set. However we might want to talk about properties of various kinds of outcomes. For example, let's say I toss a coin 3 times. There are 8 different outcomes (let's say we care about the sequence of the coin flips) and each has, if the coin is fair, a probability of 1/8. I might want to talk about the number of heads in any given outcome. That is I want to map an outcome to a number, which represents how many times heads appeared. This can be done via a function, f. So if HHH is an outcome with 3 heads, $f(HHH) = 3$. Similarly $f(HTH) = 2$ and $f(THH)$=2.

Definition 1. *(Random Variable) A Random variable is just a function $\Omega \to \mathbb{R}$ where Ω is the sample space of the probability space.*

In general the property we are talking about might not be a real number. For example if you pick a student at random from the class and ask them what their favourite song is, you could think of a random variable S. This takes a student as input and outputs the favourite song of the student. But we will mostly be looking at properties that are in the set of real numbers, so we can stick with this definition.

You may have noticed from our discussion that random variables are not random and are not variables! A random variable is a function, as we can see from the definition above. This naming is due to Francesco Cantelli, from the early 1900s.

One particularly simple, but useful random variable is what we call an **indicator random variable**. Consider a subset $A \subseteq \Omega$. $I_A(\omega) = 0$ if $\omega \notin A$ and $I_A(\omega) = 1$ if $\omega \in A$. Intuitively, I_A represents the property of being in set A.

Now that we have defined random variables, we want to be able to define the probability that a random variable takes a particular value. This can is done as follows. Say X is a random variable (Notice how we dropped the parenthesis. Actually $X()$ is a function, but we often write random variables without the arguments to the function written explicitly):

$$\Pr(X = r) = \Pr(\{\omega \in \Omega | X(\omega) = r\})$$

Let's return to our coin flipping example. Suppose we flip a coin three times. Let $X(\omega)$ be the random variable that equals the number of heads that appear when ω is the outcome. So we have

$$\begin{aligned} X(HHH) &= 3, \\ X(HHT) = X(HTH) = X(THH) &= 2, \\ X(TTH) = X(THT) = X(HHT) &= 1, \\ X(TTT) &= 0. \end{aligned}$$

Now, suppose that the coin is fair. Since we assume that each of the coin flips is mutually independent, this gives us a probability of $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ for any one of the above outcomes. Now, let's consider $\Pr(X = k)$:

k	$\{\omega \in \Omega \mid X(\omega) = k\}$	$\Pr(X = k)$
3	$\{HHH\}$	$\frac{1}{8}$
2	$\{HHT, HTH, THH\}$	$\frac{3}{8}$
1	$\{TTH, THT, HTT\}$	$\frac{3}{8}$
0	$\{TTT\}$	$\frac{1}{8}$

This looks suspiciously like something we've seen before. Here, we'll introduce the notion of Bernoulli trials.

Definition 2. *A Bernoulli Trial is a random variable B whose range is restricted to $\{0, 1\}$, where the event $\{\omega \in \Omega \mid B(\omega) = 1\}$ is the success event, and $\{\omega \in \Omega \mid B(\omega) = 0\}$ is the failure event.*

Bernoulli trials are named after Jacob Bernoulli who is also sometimes referred to as James Bernoulli in the late 1600s. There are lots of other mathematicians in the Bernoulli family but the lion's share of mathematical results named after Bernoulli are really named after this particular Bernoulli.

Bernoulli trials are a repetition of n mutually independent events. How should we interpret this in our probability framework? In other words, what do the sample space and the probability distribution look like?

If we have a probability space $(\Omega, \Pr)$ and we want to conduct n independent trials, then what we really want is to take the product of a bunch of these outcomes. So we define the space $(\Omega^n, \Pr_n)$, where Ω^n is the set of n-tuples $(a_1, a_2, \ldots, a_n)$ with $a_i \in \Omega$ and

$$\Pr_n((a_1, a_2, \ldots, a_n)) = \Pr(a_1) \Pr(a_2) \cdots \Pr(a_n).$$

Here, each position i corresponds to one of our trials and we can define the corresponding Bernoulli trial by $X_i((a_1, \ldots, a_n)) = X(a_i)$. Since we're taking the product, none of the trials in our tuple depend on each other, so they're independent. We can use this idea to prove the following.

Theorem 1. *The probability of k successes in n independent Bernoulli trials with probability of success p and probability of failure $q = 1 - p$ is*

$$\left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

Proof. For each n-tuple $A = (a_1, \ldots, a_n)$ of trials with Bernoulli random variable X, we can construct a string $w_A = b_1 b_2 \cdots b_n$ over the alphabet $\{S, F\}$ by

$$b_i = \begin{cases} S & \text{if } X(a_i) = 1 \\ F & \text{otherwise} \end{cases}$$

Then the number of n-tuples with k successes is the same as the number of binary strings of length n with $k S$'s. Recall that there are exactly $\left(\begin{array}{c} n \\ k \end{array} \right)$ of these.

Then for each n-tuple with k successes, we have $\Pr_n(A) = p^k(1-p)^{n-k}$. Putting this together, the probability of k successes is

$$\left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

□

This is called the binomial distribution, for the reason that this is what you'd get if you plugged in the $x = p$ and $y = (1 - p)$ into the Binomial Theorem. It is from this that we get the definition

$$b(k; n, p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}.$$

We can then reinterpret our results from the coin-flipping example as a Bernoulli trial. Since the probability for a fair coin flip to land on heads is $p = \frac{1}{2}$, we have

k	$b(k;n,p)$	
3	$\binom{3}{0}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0$	$\frac{1}{8}$
2	$\binom{3}{1}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1$	$\frac{3}{8}$
1	$\binom{3}{2}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^2$	$\frac{3}{8}$
0	$\binom{3}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0$	$\frac{1}{8}$

2 Expectation

By introducing random variables, we are now able to talk about the probability that some property holds true. For example, we can talk about the probability that in 3 coin flips, we get exactly 2 heads. Another natural question to ask in the coin-flipping example, is what number of heads might we expect to see on average.

Definition 3. *(Expectation) Let* $(\Omega, \Pr)$ *be a probability space and let* X *be a random variable. The expected value of* X *is*

$$E(X) = \sum_{\omega \in \Omega} X(\omega) \Pr(\omega).$$

Roughly speaking, the expectation of a random variable is the (weighted) average value for the random variable. The weights used are exactly the probabilities of the corresponding outcomes. Note that the expectation is the value you would get on average. It is **NOT** necessarily the most likely value to be obtained.

Excercise 1. *Show an example probability space and random variable where the probability that the random variable actually attains it's expected value is* 0*.*

Let's see how we might calculate the expectation of a random variable. Consider a fair six sided die. Let X be the number rolled. What is the expectation of X?

$$\begin{aligned} E(X) &= \sum_{i=1}^{6} \Pr(i) \cdot X(i) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{21}{6} = 3.5 \end{aligned}$$

But what if the dice is biased towards even numbers so that the probability of any even number is twice that of any odd number? Then the even numbers have a 2/9 probability while the odd numbers have a 1/9 probability. If Y is the random variable representing a roll of this die:

$$\begin{aligned} E(Y) &= \sum_{i=1}^{6} \Pr(i) \cdot Y(i) \\ &= 1 \cdot \frac{1}{9} + 2 \cdot \frac{2}{9} + 3 \cdot \frac{1}{9} + 4 \cdot \frac{2}{9} + 5 \cdot \frac{1}{9} + 6 \cdot \frac{2}{9} \\ &= \frac{33}{9} = 11/3 = 3.666... \end{aligned}$$

Recall that the motivation for defining random variables was so we could get away from considering the probabilities of elementary events. The definition of expectation is not very helpful in

this regard. Luckily, it is not too difficult to reformulate expectation in terms of the values that the random variable takes on.

Theorem 2. *Let* $\Pr(X = r) = \Pr(\{\omega \mid X(\omega) = r\})$ *and let* $\{r_1, \ldots, r_k\} = \text{range}(X)$. *Then*

$$E(X) = \sum_{i=1}^{k} r_i \cdot \Pr\left(X = r_i\right).$$

Proof. Recall that the range of a function X is all of the possible values that $X(\omega)$ can take over every $\omega \in \Omega$. Then we have

$$\begin{aligned}
\sum_{i=1}^{k} r_i \cdot \Pr\left(X = r_i\right) &= \sum_{i=1}^{k} r_i \cdot \Pr\left(\{\omega \in \Omega \mid X(\omega) = r_i\}\right) \\
&= \sum_{i=1}^{k} r_i \cdot \sum_{\omega \in \Omega, X(\omega) = r_i} \Pr(\omega) \\
&= \sum_{i=1}^{k} \sum_{\omega \in \Omega, X(\omega) = r_i} X(\omega) \Pr(\omega) \\
&= \sum_{\omega \in \Omega} X(\omega) \Pr(\omega) \\
&= E(X)
\end{aligned}$$

□

Just as we had a notion of independent events, we can talk about independent random variables.

Definition 4. *We say* X *and* Y *are independent random variables if* $\Pr(X = r \wedge Y = s) = \Pr(X = r) \cdot \Pr(Y = s)$.

But defining correlation is a bit different for random variables. It is done in terms of $E(X) \times E(Y)$ and $E(XY)$. However we will not get into the details of correlation between random variables in this class.

2.1 Linearity of Expectation

The following fact is very useful when trying to calculate the expectation of a sum of random variables.

Theorem 3. *(Linearity Of Expectation) Let* $(\Omega, \Pr)$ *be a probability space,* $c_1, c_2, \ldots, c_n$ *be real numbers, and* $X_1, X_2, \ldots, X_n$ *be random variables over* Ω. *Then*

$$E\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i E\left(X_i\right)$$

The proof can be obtained by just following the definitions. We will skip it here. Let's see a simple example

Suppose we have three of our dice that are biased towards even numbers. We calculated that the expected value of the roll is $\frac{11}{3}$ for each of these dice. Suppose we want to calculate the expected value of the sum of the three numbers rolled using three such dice. Without linearity of expectation we would need to list every possible tuple of rolls and calculate the probability that this particular roll occurs. But via linearity of expectation we can immediately conclude that the expected sum is

just thrice the expectation of a single such dice roll and so the answer is 11.

More formally, if X_1, X_2 and X_3 are random variables representing the roll of each biased die:

$$E[X_1 + X_2 + X_3] = E[X_1] + E[X_2] + E[X_3] = \frac{11}{3} + \frac{11}{3} + \frac{11}{3} = 11$$

Random variables that we wish to consider are often correlated in complex ways. However notice that **linearity of expectation works even when the random variables are not independent**. Let's see an example of this.

There is a dinner party where n men check their hats. The hats are mixed up during dinner, so that afterward each man receives a random hat. In particular, each man gets his own hat with probability $1/n$. What is the expected number of men who get their own hat?

Without linearity of expectation, this would be a very difficult question to answer. We might try the following. Let the random variable R be the number of men that get their own hat. We want to compute $\mathrm{E}[R]$. By the definition of expectation, we have:

$$\mathrm{E}[R] = \sum_{k=0}^{\infty} k \cdot \Pr\{R = k\}$$

Now we're in trouble, because evaluating $\Pr\{R = k\}$ is a mess and we then need to substitute this mess into a summation. Furthermore, to have any hope, we would need to fix the probability of each permutation of the hats. For example, we might assume that all permutations of hats are equally likely.

Now let's try to use linearity of expectation. As before, let the random variable R be the number of men that get their own hat. The trick is to express R as a sum of indicator variables. In particular, let R_i be an indicator for the event that the i th man gets his own hat. That is, $R_i = 1$ is the event that he gets his own hat, and $R_i = 0$ is the event that he gets the wrong hat. The number of men that get their own hat is the sum of these indicators:

$$R = R_1 + R_2 + \cdots + R_n$$

These indicator variables are not mutually independent. For example, if $n - 1$ men all get their own hats, then the last man is certain to receive his own hat. But, since we plan to use linearity of expectation, we don't have worry about independence!

Let's take the expected value of both sides of the equation above and apply linearity of expectation:

$$\begin{aligned}\mathrm{E}[R] &= \mathrm{E}\left[R_1 + R_2 + \cdots + R_n\right] \\ &= \mathrm{E}\left[R_1\right] + \mathrm{E}\left[R_2\right] + \cdots + \mathrm{E}\left[R_n\right]\end{aligned}$$

Since the R_i 's are indicator variables, $\mathrm{E}\left[R_i\right] = \Pr\left\{R_i\right\}$ and since every man is as likely to get one hat as another, this is just $1/n$. Putting all this together, we have:

$$\begin{aligned}\mathrm{E}[R] &= \mathrm{E}\left[R_1\right] + \mathrm{E}\left[R_2\right] + \cdots + \mathrm{E}\left[R_n\right] \\ &= \Pr\left\{R_1 = 1\right\} + \Pr\left\{R_2 = 1\right\} + \cdots + \Pr\left\{R_n = 1\right\} \\ &= n \cdot \frac{1}{n} = 1.\end{aligned}$$

So we should expect 1 man to get his own hat back on average!

At this point, try to work through the problem in the handout.

Consider a tennis tournament where there are 16 players. At each round the players are paired uniformly at random and the losers are eliminated. So in round 1 you have 8 matches, and only 8 players progress to round 2 where again the pairings are done uniformly at random. There are 4 matches in round 2 and so on until you have a winner. Suppose each player has exactly a half chance of beating any other player in any match. Federer and Nadal both enter the tournament and you really want to see them play against each other. You are willing to go to every match in the tournament to see these two play against each other. What is the chance that the two will face off at some point in the tournament?

We might initially try to consider various ways in which the two players could end up meeting, and individually calculate their probabilities, but this soon gets messy. Instead we can proceed in a different way so as to exploit the power of the linearity of expectations.

Let's number the players from 1 to 16. Suppose Federer is player 1 and Nadal player 2. We want the chance that player 1 plays player 2 at some point. But notice that all the rules of the tournament are symmetric with respect to the players. That is the chance that A and B play each other at some point any given point is no different than the chance C and D play each other at that point. Let $I_{i,j}, i < j$ be the indicator random variable, indicating whether i and j play each other at any point in the tournament. What we would like to calculate is $I_{1,2}$.

Notice that

$$E[I_{1,2}] = 1 \cdot \Pr(i \text{ plays } j \text{ at some point}) + 0 \cdot \Pr(i \text{ does not play } j) = \Pr(i \text{ plays } j \text{ at some point})$$

So we just need to calculate $E[I_{1,2}]$. There are $\binom{16}{2}$ possible pairs of players and we know the probability of any two players playing each other is the same. Let's call this probability p. Then $E[I_{1,2}] = p$ and also

$$E[\sum_{1 \le i < j \le 16} I_{i,j}] = \sum_{1 \le i < j \le 16} E[I_{i,j}] = \binom{16}{2} p$$

But We know what $\sum_{1 \le i < j \le 16} I_{i,j}$ is. It is just the number of matches that occurs. Each match eliminates 1 player, so the number of matches is 15. Thus $E[\sum_{1 \le i < j \le 16} I_{i,j}] = 15$. So $\binom{16}{2} p = 15$ or $p = 1/8$.

Notice that the $I_{i,j}$ random variables are not independent, if I know player 1 played players 3, 4, 5 and 6 then the chance that he played 2 is exactly 0, since each player plays only 4 matches at the most.

2.2 The 3-SAT problem and a Monte Carlo algorithm

The 3-satisfiability problem (3-SAT) is one of the most famous "difficult" problems in computer science. The problem asks, when given a propositional formula in conjunctive normal form where each clause has three variables, whether there is an assignment of variables that makes the formula evaluate to true. Conjunctive normal form means that the proposition has been written as the AND of a bunch of terms, each term being the OR of some variables or their negations. Here, a clause is three variables or their negations joined together by $\vee$. Each clause is joined by $\wedge$. For example,

$$(x_1 \vee \neg x_2 \vee x_4) \wedge (x_2 \vee x_3 \vee \neg x_4) \wedge (\neg x_1 \vee \neg x_3 \vee x_5)$$

One reason we might care about this problem is that this compound proposition in CNF form can be thought about as encoding a bunch of constraints and we want as many as possible to be satisfied. For example say you have a bunch of classes you want to schedule and for each you want

one of three requirements (say related to class size, time, location etc) to be met. This can be mapped onto an instance of the 3-SAT problem. The satisfiability in general can be used to model a large class of constraint satisfaction problems.

One variant of this problem asks for an assignment that makes as many clauses true as possible. This is difficult, because if there are n variables, we have to look through as many as 2^n different assignments to see which one gives us the maximum number of satisfied clauses.

However, what if we just randomly guessed an assignment? Suppose we flip a coin independently for each variable x_i, setting it to True with probability $\frac{1}{2}$. How many clauses should we expect to satisfy? Our claim is the following.

Theorem 4. *Let φ be a formula in 3-CNF with k clauses. we should expect that $\frac{7}{8}$ of them will be satisfied.*

This might be surprising, just randomly assigning values allows us to satisfy a good fraction of the clauses.

Proof. Let's name our clauses $C_1, \ldots, C_k$. Define the random variable Z_j by

$$Z_j = \begin{cases} 1 & \text{if clause } C_j \text{ is satisfied} \\ 0 & \text{otherwise} \end{cases}$$

Then we define the random variable Z to be the number of clauses that have been satisfied. That is, $Z = Z_1 + \cdots + Z_k$. We can then compute $E(Z)$ by figuring out $E(Z_j)$ for a single clause. But what is $E(Z_j)$?

Let's consider the possible assignments we can make to a clause $(x \vee y \vee z)$. There are three variables (or their negations) and each one can take on two values, so there are as many as 8 possible assignments. However, we observe that the clause is satisfied as long as one of the terms in it evaluates to True. In other words, in order for the clause not to be satisfied, all three terms must evaluate to False. There is only one possible assignment that does this, with probability $\frac{1}{8}$, which means that the probability that our clause is satisfied is $1 - \frac{1}{8} = \frac{7}{8}$.

So $E(Z_j) = \frac{7}{8}$ for all j. This gives us

$$E(Z) = \sum_{j=1}^{k} E(Z_j) = \sum_{j=1}^{k} \frac{7}{8} = \frac{7}{8}k$$

□

This is a nice result: it says that if we have a random assignment, we can expect to satisfy 7/8 of the clauses. Now, notice that we are not guaranteed this number in any particular try—we may end up satisfying fewer. However, we could also end up satisfying more. But we can actually conclude that there has to exist an assignment of variables that satisfies at least 7/8 of the clauses.

This is another result that uses the probabilistic method—we showed that there is a nonzero probability that such an assignment exists, because if there was no way to satisfy at least 7/8 of the clauses, then the average number of satisfied clauses couldn't be 7/8 of the total.

More formally:

Proposition 1. *If X is a random variable on a finite probability space such that $E[X] = v$, then there is some event ω such that $X(\omega) \geq v$.*

Convince yourself of this. But though we now know there is at least 1 assignment that satisfies at least 7/8 of the clauses, we don't have any way of finding it easily, which doesn't really help us solve the problem. Or does it?

Suppose we extend our previous strategy. Instead of just randomly assigning variables and taking the result, what if we kept on coming up with random assignments until we get an assignment that satisfies at least 7/8 of the clauses? Suppose we knew there is at least a p chance that a random assignment . How long should we expect to have to keep trying random assignments? The answer is provided by the following theorem. We will skip over the proof here, since it involves some calculations with infinite series, but you can try to work it out on your own. The result is also somewhat intuitive.

Theorem 5. *Let X be a Bernoulli trial with probability of success p. The expected number of trials before the first success is $1/p$.*

For example, if there is a 10% chance of a lightbulb being defective, the expected number of bulbs we need to buy before we encounter a defective one is 10. Applying this to our 3-SAT problem, if I can show that there is at least a p chance of a random assignment satisfying at least 7/8 of the clauses, then, in expectation $1/p$ random tries should suffice.

Theorem 6. *Let φ be a formula in $3-$CNF with k clauses. The probability that a random assignment satisfies at least $\frac{7}{8}k$ clauses is at least $\frac{1}{8k}$.*

Proof. Let p_j denote the probability that exactly j clauses are satisfied by a random assignment. We want to compute the probability $p = \sum_{j \geq \frac{7}{8}k} p_j$ that at least $\frac{7}{8}k$ clauses are satisfied. By definition (see the proof of Thm. 4), we have

$$E(Z) = \frac{7}{8}k = \sum_{j=0}^{\infty} j p_j$$

We split this sum into two:

$$\sum_{j=0}^{\infty} j p_j = \sum_{j < \frac{7}{8}k} j p_j + \sum_{j \geq \frac{7}{8}k} j p_j$$

We note in the first sum that j ranges from 0 to just under $\frac{7}{8}k$, which gives us

$$\sum_{j < \frac{7}{8}k} j p_j \leq \left(\frac{7}{8}k - \frac{1}{8}\right) \sum_{j < \frac{7}{8}k} p_j.$$

Similarly, the second sum has j ranging from $\frac{7}{8}k$ to k, the total number of clauses. This gives us

$$\sum_{j \geq \frac{7}{8}k} j p_j \leq k \sum_{j \geq \frac{7}{8}k} p_j$$

Together, we have

$$\begin{aligned}
\sum_{j=0}^{\infty} j p_j &= \sum_{j < \frac{7}{8}k} j p_j + \sum_{j \geq \frac{7}{8}k} j p_j \\
&\leq \left(\frac{7}{8}k - \frac{1}{8}\right) \sum_{j < \frac{7}{8}k} p_j + k \sum_{j \geq \frac{7}{8}k} p_j \\
&= \left(\frac{7}{8}k - \frac{1}{8}\right) \cdot (1-p) + kp \\
&\leq \left(\frac{7}{8}k - \frac{1}{8}\right) + kp
\end{aligned}$$

To summarize, we have

$$\frac{7}{8}k \leq \left(\frac{7}{8}k - \frac{1}{8}\right) + kp$$

We rearrange this to solve for p :

$$\begin{aligned}
\left(\frac{7}{8}k - \frac{1}{8}\right) + kp &\geq \frac{7}{8}k \\
kp &\geq \frac{7}{8}k - \left(\frac{7}{8}k - \frac{1}{8}\right) \\
kp &\geq \frac{1}{8} \\
p &\geq \frac{1}{8k}
\end{aligned}$$

□

Using this result, we can conclude that we expect the number of assignments we need to try (i.e. the number of trials before a success) is $8k$, which is linear in the size of a 3-CNF formula. This leads to a straightforward Monte Carlo algorithm. Monte Carlo algorithms are algorithms that have some chance of producing an "incorrect" answer (in this case, an assignment that satisfies fewer than $\frac{7}{8}$-th of the clauses). The key insight that makes Monte Carlo algorithms work is that we can simply retry our algorithm until we have a good enough answer, and we can prove a bound on how many times we expect to re-run it.

3 Markov's Inequality: A Concentration Inequality

While we certainly cannot say that the expected value of a random variable is the one we should expect to turn up very often, we can actually bound the probability that a random variable has a value that is far from it's expectation.

let's say we have 50 students in a class and they write a test with a maximum possible score of 100 points and a minimum possible score of 0 points. Suppose the average score is 60/100. How many students could possibly have a score of 90/100 or more?

Well if the average is 60, then the total score of all students is $50 \times 60 = 3000$. Even if all other students had a score of 0 and some k students had a score of 90/100, then the total score would be at least $90k$. Once $k = 34$, then the total score would already be at least $90 \times 34 = 3060$, which is too big. Thus at most 33 students could have scored 90 or above.

This is exactly the reasoning which, when generalized, yields Markov's theorem.

Theorem 7. *If X is a nonnegative random variable and $a > 0$, then the probability that X is at least a is at most the expectation of X divided by a.*

$$\Pr(X \geq a) \leq \frac{E(X)}{a}$$

Notice that if we apply this theorem to our previous example, we could let X be the random variable denoting the score of a uniformly randomly drawn student. Then $E(X) = 60$. We obtain

$$\Pr(X \geq 90) \leq \frac{E(X)}{90} = \frac{60}{90} = 2/3$$

If k students out of 50 score 90 or above, then the chance a uniformly randomly drawn student scores at least 90 is $k/50$. Thus

$$\frac{k}{50} \leq \frac{2}{3}$$

Or $k \leq 100/3$, so k can at most be 33. This is the same answer we obtained by our intuitive calculation. In this case Markovs theorem for gives us a good bound on the probability of the random variable being much larger than the expectation. However, in general, the bounds provided by Markov's theorem can be quite weak. Try to come up with an example where this is the case.

The proof of Markov's inequality actually follows our reasoning in the example about the students and the test quite closely. You can begin by writing out the definition of the expectation and then show that if the theorem were not true, then the expectation would be too large.

Excercise 2. *Prove Markov's theorem.*

4 Variance

Markov's inequality provides a one-sided bound on deviation from the expectation. That is , it tells us that a random variable is not all that likely to be hugely bigger than it's expectation. But what if we want to know how certain we can be that a random variable will not be much larger or smaller than it's expectation. That is, we want a two-sided bound, which is something that would tell us how likely a random variable is to be concentrated around it's expectation. Chebyshev's theorem, a variant of Markov's theorem, allows us to do this. However we will not study this theorem in this class. Instead we will conclude today's lecture by defining a notion that measures how spread out, in either direction a random variable is from its mean.

Definition 5. *(Variance) The variance of a random variable X is denoted by Var(X) and is defined as $E((X - E(X))^2)$.*

This might make little intuitive sense at first. What even is $E((X - E(X))^2)$? But we can think about it as follows. We, in some sense, expect X to have values around $E(X)$. We now want to measure how far away X is, on average, from $E(X)$. $(X - E(X))^2$ is a good way to measure this. If X is much lower or higher than X, then it will be large, while if X is near $E(X)$ it will be small. So what we want to measure is the average value of $(X-E(X))^2$, which is $E((X-E(X))^2)$.

Consider the following two games.

Game 1:A fair coin is tossed. If it is heads you are given 1 dollar, but if it is tails you are given 0 dollars.

Game 2: A fair coin is tossed. If it is heads you are given 101 dollars, but if it is tails you must pay 100 dollars.

The expected profit in both games is the same. Examine the random variable R that represents your profit from the game. Check that the expected profit is 0.5 dollar. Now calculate the variance of R in both cases. Intuitively, it will be much higher for game 2.

For game 1 we have:

$$\begin{aligned}
\mathrm{Var}(R) &= E((R - E(R))^2) \\
&= \Pr(R = 1)(1 - E(R))^2 + \Pr(R = 0)(0 - E(R))^2 \\
&= 0.5 \times (1 - 0.5)^2 + 0.5 \times (0 - 0.5)^2 \\
&= 0.5^3 + 0.5^3 \\
&= 2 \times 0.5^3 = 0.25
\end{aligned}$$

For game 2 we have:

$$\begin{aligned}
\mathrm{Var}(R) &= E((R - E(R))^2) \\
&= \Pr(R = 101)(101 - E(R))^2 + \Pr(R = -100)(-100 - E(R))^2 \\
&= 0.5 \times (101 - 0.5)^2 + 0.5 \times (-100 - 0.5)^2 \\
&= 0.5 \times 100.5^2 + 0.5 \times 100.5^2 \\
&= 100.5^2
\end{aligned}$$

What about the variance of a sum of two random variables $X + Y$? Unlike the expectation, where we could always just add the expectations, we can in general only do this for the variance if the random variables are uncorrelated. In particular, for independent random variables X and Y, we have:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$