Lecture 1: 205 recap, sum rule, and the union bound

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapters 1,3,4,5, and 15 of the Math for CS book

1 What is the course about?

The course is about counting and probability basics. These tools are helpful to know as a computer scientist, no matter which subfield of CS you get into. I will post some reading material which gives examples of applications of counting and probability in CS. To get an idea about the topics that will be covered in this course check out the syllabus PDF on Sakai.

2 Lecture notes

Note that the purpose of the lecture notes is to supplement the lectures, and not to replace them! You will find that the lecture notes are a barebone sketch of what actually happened in the lecture. My main motivation to type these out is

1. to let students who missed the lecture get an idea as to what was covered in class (and how awesome it was, and why they shouldn't miss a class again), and

2. to stimulate the associative memory of students who did attend the class, serving as a sort of a refresher.

The notes might have typos/minor mistakes (hopefully not major stuff), so please feel free to point them out.

3 How to do well in this course?

First and foremost, attend the lectures and recitations, and bug your instructor and TAs with any doubts/questions you have. Second, practice, practice, and practice! The *Mathematics for Computer Science* book has a whole bunch of practice problems. Additionally, you'll have enough problems to ponder over between your HWs, Quizzes, and practice problem sets.

Oh, and we have office hours (4 of them in a week!) so make sure you drop by and make use of them! You can also discuss problems (not HW problems though) on Piazza!

4 Getting ready: a 205 recap

We will be using a lot of 205 concepts in this course so the first thing you want to do is to brush up on those concepts if you feel they are getting rusty! In this lecture, we will try to do a crash course on the essential 205 concepts. You'll also get a glimpse into future topics involving counting!

4.1 Set theory

Without getting into super-formal definitions, let's just say that a set is a *collection* of objects. See if you can remember what these mean:

- Commonly used sets: $\mathbb{R}, \mathbb{Z}, \mathbb{N}, \mathbb{Q}, \mathbb{C}$, etc.
- Set operations: $X \cap Y, X \cup Y, X \setminus Y, X \times Y, |X|, 2^X, X^c$.

Here are some exercises to test you:

Question . Is it always the case that for two sets X and Y, $X \setminus Y = Y \setminus X$? If not, give a counterexample!

If X is a set then 2^X or $powerset(X)$ is called the *power set* of X and it is the set of all subsets of X. For example, if $X = \{1, 2\}$, then $2^X = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$.

Often, we will use the *set builder* notation to define a set, e.g. $X = \{n \in \mathbb{N} | \text{ n is a prime}\}$. Here are some more problems to test yourself:

Question . Let $A = \{n \in \mathbb{N} : \ n^2 < 49\}$ and $B = \{n \in \mathbb{N} : \text{ n is a prime}\}$. Find $A \cap B, A \cup B, A \setminus B, A \times B$.

$A \times B$ is the cartesian product and is the set of all pairs (a, b) where $a \in A, b \in B$. Notice that these are ordered pairs, unlike the set $\{a, b\}$ which has no innate order. One can extend this to the n-fold cartesian product: $A_1 \times A_2 \times \ldots \times A_n = \{(a_1, a_2, \ldots, a_n) | \ a_1 \in A_1, a_2 \in A_2, \ldots, a_n \in A_n\}$.

The complement of a set A denoted by A^c only makes sense if you know what *universe* A is inside. For example, the set $1, 2$ could be inside the universe $\mathbb{N}$ or the universe $\mathbb{Z}$. Usually it is clear from context what the universe is. For example, if $A = \{1, 2, 3\}$ is a subset of the universal set $S = \{1, 2, 3, 4, 5\}$, then A^c, the complement, is $\{4, 5\}$.

Sometimes it's useful to represent sets visually. This can be done using *Venn diagrams.* In a Venn diagram, every set is represented by a circle, and the intersection between two sets can be represented by showing the circles corresponding to the sets as overlapping. On the other hand if two sets are disjoint, then one can represent that as two non-overlapping/non-intersecting circles. It's useful to bound the sets inside a large rectangle that represents the universe inside which the sets live.

Now let's test your knowledge on the above:

Question . A school has 200 students. 25 of them like both CS and Math, 100 of them like CS, and 50 of them like Math. How many of the students like neither of the two subjects?

4.2 Functions

A function from set A to set B, denoted by $f : A \rightarrow B$, "associates" with EVERY $a \in A$ some elememt $b \in B$. Read that again: for EVERY $a \in A$. A is called the domain, and B the codomain. We say that f *maps* the elements in A to elements in B. Note again, it maps EVERY element in A to something in B – there cannot be elements in A that f does not map to an element in B. Of course, two elements in A, say a_1 and a_2 could get mapped to the same $b \in B$. That's totally fine!

When we write $f(a) = b$ for some $a \in A$ and $b \in B$ we mean that a if mapped by f to b. For example, let us define the function f that maps $\mathbb{R}$ to $\mathbb{R}$ (i.e., $f : \mathbb{R} \rightarrow \mathbb{R}$) as follows:

$$f(x) = 4 \cdot x.$$

This means that the real number x is mapped to the real number $4 \cdot x$. 2 is mapped to 4, 4 is mapped to 16, etc. There are different "types" of functions:

1. Injective functions: A functon $f : A \rightarrow B$ is injective if for $a_1, a_2 \in A$ (where $a_1 \neq a_2$, i.e. they are not the same element), $f(a_1) \neq f(a_2)$. This means no two elements in A get mapped to the same $b \in B$ by f. See if you can prove this: if $f : A \rightarrow B$ is injective then it MUST be the case that $|B| \geq |A|$. Example $f(x) = 4 \cdot x$ where $f : \mathbb{R} \rightarrow \mathbb{R}$.

2. Surjective functions: A function $f : A \rightarrow B$ is surjective if for every $b \in B$ there is some $a \in A$ such that $f(a) = b$, i.e. some $a \in A$ is mapped to b by f. No $b \in B$ is left alone!
   You can check that if f is surjective then it MUST be the case that $|A| \geq |B|$. Examples of functions that are surjective: $f : \mathbb{Z} \rightarrow \mathbb{N}$, $f(x) = |x|$ (absolutve value of x).

3. Bijective functions: If a function is both surjective and injective, then it is a bijective function! If $f : A \rightarrow B$ is bijective, what can you say about the relation between $|A|$ and $|B|$?

Sometimes we will use the names *injection, surjection, and bijection* instead of injective function, surjective function, and bijective function.

4.3 Propositions, predicates, axioms, rules of logic, and proofs

Whenever we "do" mathematics, we implicitly deal with the following "entities" from logic:

1. *Proposition*: A proposition is a statement or an assertion that may be true of false (depending on our assumptions). For example, *Every number n has a number larger than it* is a proposition (and it's a true proposition, assuming we are dealing with, say, real numbers), and 4 *is a prime number* is also a proposition, although it's a false one ($4 = 2 \times 2$).

2. *Axioms*: These are our basic assumptions that we make about mathematical objects: for example, if A is a subset of $\{1, 2, \ldots, n\}$, either the number 1 is in A or it's not in A — it cannot both be in and not be in A, or the fact that adding any number to zero gives you back the same number. These are things we just assume and don't try to prove — these are examples of axioms!

3. *Predicate*: Simply put, a predicate is a proposition that has variables in it. For example, X *is divisible by* 2 is a predicate. We cannot say whether it's true or false unless we specify

what X is, so in some ways it's like a function — it's "output" or "value" depends on what "input" you feed it! If $X = 4$, then it becomes true, but if $X = 3$, it's false!

4. *Quantifiers*: You can convert a predicate into a proposition by adding quantifiers. For example, the above predicate becomes a proposition if I add the "for all" quantifier: $\forall X \in \mathbb{N}$, *X is divisible by 2*. Is this proposition true or false?
   You can also add the "there exists" quantifier: $\exists X \in \mathbb{N}$, *X is divisible by* 2. (Is this true or false?).

5. *Rules of logic*: These are rules that we use to go from a set of propositions and axioms to a new proposition. For example, if I have that the proposition $P \implies Q$ is true and have P as an axiom, then together they imply that the proposition Q is true. I wouldn't list all the rules of logic here. You can find a comprehensive list and discussion in Rosen's book (Discrete Math and Its Applications) or in the Mathematics for Computer Science book (See the reading suggestions on Sakai/the course webpage).

6. *Proofs*: A proof starts out with the axioms and other assumptions, and at every step obtains a new proposition from the axioms, assumptions, and previously obtained propositions, using the rules of logic.

4.4 A glimpse into the zoo of different types of proofs

There are different styles of proof that you would have seen in 205. Here are some:

1. *Direct proof*: If you want to prove $P \implies Q$, the simplest way to do so is to assume P is true and then derive/prove Q from P. For example, if you want to prove the following:

   $$1 \leq x \leq 2 \implies x^2 - 3x + 2 \leq 0,$$

   what you can do is to assume that x is between 1 and 2 (inclusive), observe that $x^2 - 3x + 2 = (x-2)(x-1)$ and that, under the assumption, $(x-2) \leq 0$ and $(x-1) \geq 0$, and thus $x^2 - 3x + 2 = (x-2)(x-1) \leq 0$.

2. *Proof by contrapositive*: Instead of directly proving that $P \implies Q$, you prove the contrapositive, i.e, $\neg Q \implies \neg P$, which is a logically equivalent statement. For example, if you want to show that if r is irrational then so is $\sqrt{r}$, you can show the contrapositive:

   If $\sqrt{r}$ is rational, then so is r.

   If $\sqrt{r}$ is rational, then $\sqrt{r} = a/b$ (for positive integers a and b), and so $r = a^2/b^2$, which means that r is rational.

3. *Proof using a case analysis*: Say you want to show that the function $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) = |x+2| - |x-3|$ always takes a value in the interval $[-5, 5]$. Recall that $|x| = x$ if $x \geq 0$, and $|x| = -x$ if $x < 0$. Thus, it makes sense to analyze the behavior of $f(x)$ for different cases depending on the values of x. For example, in the first case we will look at all values of x where $(x+2) < 0$ but $(x+3) \geq 0$. The only value of x that satisfies these conditions is $x = -3$, and in this case $f(x)$ becomes $-x - 2 - x + 3 = 1$, which is in the interval $[-5, 5]$. One can similarly look at other cases and prove the statement for each of the cases.

4. *Proof by contradiction*: If you want to show that a statement/proposition P is true, you assume for the sake of contradiction that P is false, and use that to arrive at some contradiction (a contradiction to one of your axioms or the assumption you started out with). For example, if you want to prove that $\sqrt{2}$ is irrational, you can prove this using proof by contradiction: assume $\sqrt{2}$ is rational and thus $\sqrt{2} = a/b$ such that a and b are positive integers that don't share any common divisors. Then $a^2 = 2 \cdot b^2$. But this means that a^2 is a perfect square that is also even, and so a itself must be even (why?). This means 4 divides a^2, and thus 4 also divides $2 \cdot b^2$ (since $a^2 = 2 \cdot b^2$). This would mean that b^2 is even and thus b is even. So we have shown that both a and b are even, which contradicts the assumption that a and b have no common divisors. Thus, $\sqrt{2}$ cannot be rational.

5. *Proof by weak induction*: Induction is useful for proving statements/propositions of the form

$$\forall n \in \mathbb{N},\ P(n) \text{ is true},$$

where $P(n)$ is a predicate on natural numbers. The idea is to first prove that $P(0)$ is true (this is called the base case. [1]), and then showing that for every $n \geq 0$, $P(n)$ implies $P(n+1)$ (this is the induction step).
For example, to show that the sum of the numbers from 0 to n is $n(n+1)/2$, for all $n \geq 0$, we first observe that the base case ($n = 0$) is trivially true. For the induction step, let n some arbitrary integer greater than 0, and suppose that the sum of 1 to n is $n(n+1)/2$. We will use this to show that the sum of numbers from 1 to $n+1$ is $(n+1)(n+2)/2$, thereby finishing the proof. To see this, note that $1+2+\ldots+n+n+1 = (1+2+\ldots+n)+n+1 = (n(n+1)/2) + (n+1)$, where the last step follows from the inductive assumption. Thus, $1+2+\ldots+n+1 = (n(n+1)/2) + (n+1) = (n+1)(n+2)/2$, and this finishes the proof.

6. *Proof by strong induction*: Sometimes proofs using weak induction can get pretty nasty, and in such cases strong induction comes in handy. Strong induction works exactly the same as way as weak induction except the induction step: in strong induction we have to prove that for all $n \in \mathbb{N}$,$P(0) \wedge P(1) \wedge \ldots P(n) \implies P(n+1)$, i.e. we can now assume not only $P(n)$, but all of $P(0), P(1), P(2), \ldots, P(n)$ to prove $P(n+1)$. This can make the job much easier in many cases.
Consider the following problem: prove that every integer $n \geq 1$ can be written as a sum of distinct powers of two. The base case $n = 1$ is pretty easy since $1 = 2^0$. Let's prove the induction step using strong induction. Let n be arbitrary, and suppose $P(1), P(2), \ldots, P(n)$ are all true. We want to prove that $P(n+1)$ is true. Let us look at the case when $n+1$ is odd. Using the inductive assumption we know $P(n)$ is true, i.e. n can be written as a sum of distinct powers of two: $n = 2^{a_1} + \ldots + 2^{a_k}$, for some k distinct powers $a_1, \ldots, a_k$. Note that 2^0 does not occur in the sum-of-distinct-powers-of-two expression for n since n is even (why?). Thus, we can write $n+1$ as $1+n = 2^0 + (2^{a_1} + \ldots + 2^{a_k})$. Now consider the case when $n+1$ is even. Unlike the odd case, in this case, 2^0 can occur in the sum-of-distinct-powers-of-two expression for n, and so we cannot simply add 2^0 to that expression. Instead, we will look at the expression for $(n+1)/2$ (since it is an integer). Because we are using strong induction, we know that $P((n+1)/2)$ is true, and thus we have a sum of powers of two representation for $(n+1)/2$. Say the representation is $(n+1)/2 = 2^{b_1} + \ldots + 2^{b_\ell}$, where the b_is are distinct

[1] The base case need not always be $n = 0$. It really depends on what statement you are trying to prove. For example, sometimes, the statement might only assert something about integers $n \geq 2$, in which case your base case will be $n = 2$.

numbers. But this means that $n + 1 = 2^{b_1+1} + \ldots + 2^{b_\ell+1}$ (this is just multiplying both sides by 2 in the previous expression). This finishes the proof.

This finishes our review of some important 205 concepts. Be sure to practice and study a bit more on your own!

5 The sum rule

In counting, the goal is to find $|X|$ for some set X. Of course, if you knew X “explicitly” you could just “count” how many elements it has. However, in most situations, you only have an “implicit” description of X (and this will become clearer in the next lecture when we see tons of examples of counting problems), and this makes counting the number of elements in X challenging.

The first basic rule we have for counting is the *sum rule*: suppose A and B are disjoint sets, then

$$|A \cup B| = |A| + |B|.$$

This commonsensical rule has far-reaching applications (as we shall see next time). When A and B are not disjoint, we know that $|A \cup B| = |A| + |B| - |A \cap B| \leq |A| + |B|$. The inequality $|A \cup B| \leq |A| + |B|$ is called the *union bound*, and we shall see powerful applications of it later on in the course.

Lecture 2: Sum rule, partition method, difference method, bijection method, product rules

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 15 of the Math for CS book.

1 Sum rule and the partition method

Recall the sum rule from the first lecture: if A and B are disjoint sets, then

$$|A \cup B| = |A| + |B|.$$

We can generalize this to any finite collection of disjoint sets: if $A_1, \ldots, A_n$ are disjoint, then

$$|A_1 \cup \ldots \cup A_n| = \sum_{i=1}^{n} |A_i|.$$

Suppose you are asked to compute the cardinality of a set A. One way to do so is to partition A into finitely many disjoint parts $A_1, \ldots, A_n$ such that together those parts cover A (that is what a partition means in the first place, so we are being redundant. I just wanted to make sure you understand what we mean by partitioning A.), compute the cardinality of each of the parts separately, and then recover the cardinality of A using the sum rule. Let's see this in action using a problem:

Question . Suppose we roll two die, a black one and a white one, together. By a configuration or outcome we mean the numbers that show up on the die. We can specify an outcome using a tuple (b, w) where the first component is the number that shows up on the black dice, and the second component is the number that shows up on the white dice. If A is the set of all outcomes, what is $|A|$?

To compute the cardinality of the set of all possible outcomes, we break A into 6 parts: $A_1, \ldots, A_6$, where A_i is the set of all outcomes where the number on the black dice is equal to i. Notice that for each i we have that $|A_i| = 6$ (why?). Moreover, for every $i \neq j$, we have that $A_i \cap A_j = \emptyset$ (why?), and that $\bigcup_{i=1}^{6} A_i = A$. Thus, using the sum rule, $|A| = \sum_{i=1}^{6} |A_i| = 36$.

Use the same idea as above to solve the following problems:

Question . Suppose the setup is exactly as in the previous question. Compute the cardinality of each of the following sets:

- The set of all outcomes where the black and white dice end up with different numbers on them.
- The set of all outcomes where the number on the white dice is strictly greater than that on the black dice.

2 Difference method

Suppose we are dealing with a set A whose cardinality seems to be hard or cumbersome to compute using the partition method. A way to deal with such situations is the *difference method.* The idea is to find a larger set S such that $A \subset S$, such that the cardinality of S and $S \setminus A$ are easy to compute. Once you know $|S|$ and $|S \setminus A|$, using the sum rule, we know that

$$|S| = |A| + |S \setminus A| \implies |A| = |S| - |S \setminus A|.$$

Let us use this method to quickly compute the cardinality of the first set in the previous question (black and white dice have different numbers). Let A be the set of all outcomes where the two die have different numbers. Let S be the set of all possible outcomes. We know that $A \subset S$, and $|S| = 36$. Notice that $S \setminus A =$ all outcomes where the two die have the same number on them, and thus $|S \setminus A| = 6$. This then easily gives us $|A| = 30$! Cool, isn't it? Now use this method to solve the following problem:

Question . Find all 10-bit integers that have at least 2 ones in their binary representation.

(If you don't know what a 10-bit number is, or what a binary representation is, it might be time to ask your TA, or brush up some 205 concepts!)

3 Bijection method

Here is yet another method that's useful for computing $|A|$ when it's cumbersome to do so directly. Suppose you could find a set B such that i) it's easy to count and ii) there is a bijection $f : A \to B$, then we know from our review in the last lecture that $|A| = |B|$. This is called the *bijection method.* Let's use this method to solve find the cardinality of the set of all outcomes in the two-dice experiment where the white dice ends up with a strictly larger number than the black dice. Let's denote this set by $A_{w>b}$, and let's denote the set of outcomes where they end up with same numbers by $A_{w=b}$, and the set where the number on the black dice is strictly greater than the one on the white dice by $A_{b>w}$.

First notice that the three sets are disjoint, and account for all possible outcomes. Thus, by the sum rule, $|A| = |A_{w>b}| + |A_{w=b}| + |A_{b>w}|$, where A is the set of all possible outcomes. Recall that $|A| = 36$ and $|A_{w=b}| = 6$. Thus, we get $|A_{w>b}| + |A_{b>w}| = 30$.
Here the cool part now: we can set up a bijection between $A_{w>b}$ and $A_{b>w}$. Given a configuration/outcome in set $A_{w>b}$, we can map it to a unique configuration/outcome in the set $A_{b>w}$ by swapping the numbers on the white and black dice. For example, the outcome in the set $A_{w>b}$ where white has 5 and black has 1 can be mapped to the outcome in $A_{b>w}$ where we swap the two numbers, i.e. black has 5 and white has 1. Convince yourself that this is indeed a bijection between the two sets. This means that $|A_{w>b}| = |A_{b>w}|$, and thus, $|A_{w>b}| = 15$.

Question . How many subsets of $\{1, 2, \ldots, 101\}$ are there that have an odd number of elements?

(We shall see very soon that the total number of possible subsets of a set S of size n is 2^n.) To solve this problem, one can set up a bijection between the set of all odd size subsets of $\{1, 2, \ldots, 101\}$

and the set of all even subsets of $\{1, 2, \ldots, 101\}$ in the following way: given an odd size subset S, we map it to $\{1, 2, \ldots, 101\} \setminus S$, i.e. its complement inside $\{1, 2, \ldots, 101\}$. Why is this map a bijection? What's the final answert then?

4 Product rule

In its simplest form, the product rule looks kind of obvious:

$$|A \times B| = |A| \cdot |B|,$$

where A and B could be two arbitrary sets (not necessarily disjoint). I claim that you can derive the product rule using the partition method. How? Assume that $|A| = k$, and $A = \{a_1, \ldots, a_k\}$, and partition $A \times B$ into k parts A_i for $1 \leq i \leq k$, where A_i is all the elements of $A \times B$ which have a_i as their first component (in the tuple). Now complete the proof.

Of course, the product rule can be generalized to n sets $A_1, \ldots, A_n$:

$$|A_1 \times A_2 \ldots \times A_n| = \prod_{i=1}^{n} |A_i|.$$

(the symbol $\prod$ is the equilvalent of $\sum$ for products.)

Whenever you see that you are dealing with a set that contains sequences, tuples, or outcomes from an experiment repeated multiples times, you might want to consider using the product rule for computing its cardinality (of course, this is not a hard and fast rule, just a rule-of-thumb. We will see cases where this will not work in a straightforward manner).

For example, going back to the black and white die experiment, we can estimate the total number of outcomes pretty easily now: if A is the set of outcomes for the white dice, and B the set of outcomes for the black dice, then the total number of outcomes is basically $|A \times B|$ (why?), and therefore, using the product rule, is equal to $|A| \cdot |B| = 36$.

Question . A valid code must satisfy the following conditions:

- It has total length four.
- The first two characters are digits between 0 and 9.
- The third character is an upper case or lower case letter (they are considered to be different possibilities, i.e. $a \neq A$).
- The last character must be one of the following: #,*, or @.

How many valid codes are there?

This can be viewed as finding $|A \times A \times B \times C|$, where A contains the digits $0 - 9$, B contains the upper and lower case letters (52 in all), and C contains three special characters specified above. What's the answer?

Question . A valid password must satisfy the following conditions:

- It has total length six.
- Each character is an uppercase letter or a digit.
- The password must have at least one digit.

How many valid passwords are there?

Hint: Find all possible strings of length 6 that can be made from uppercase letters and digits, find all length six strings that follow the above rules except the last one, and use difference method to finish the problem.

5 Generalized product rule

Suppose you have 100 chairs arranged in a row. In how many ways can you seat 100 people? In other words, how many arrangements of people are there. For the first chair, any of the 100 people can be seated there. Once you fix the first choice, for the second chair there are 99 choices. Having fixed the first two choices, there are 98 choices for the third chair, and so on. It turns out that the answer is: $100 \times 99 \times 98 \times \ldots \times 2 \times 1$. This is denoted by 100! (called "hundred factorial"). In general,

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1,$$

is the number of possible arrangements/permutations of n distinct objects.

The counting we did above falls under the *generalized product rule*: suppose we are trying to find the number of length n sequences such that there are P_1 choices for the first element in the sequence, P_2 choices for the second element in the sequence for every fixed choice of the first element, P_3 choices for the third element for every fixed choice of the first and second element, ..., and P_n choices for the n^{th} element in the sequence for every fixed choice of the first $n-1$ elements, then the total number of such sequences is $P_1 \cdot P_2 \cdot \ldots \cdot P_n$.

Note that this is different from the product rule: in the product rule, the choices for each element in the sequence were in some sense "independent" of each other. For example, in the question on the number of valid codes, each character in the code could vary independently in its respective set (the first character could be any digit, independent from the second character, and so on), however, in the generalized product rule, we do not have such "independence of choice" for different elements in the sequence. Going back to the 100 chairs example, note that each chair could have any of the 100 people seated in it, however, these choices cannot vary independently of each other: if person 1 sits in chair 1, then no other chair can seat person 1. It's important to be able to distinguish between the two versions of the product rule.

Question . In how many ways can we place two rooks, a black one and a white one, on an 8×8 chessboard so that they cannot attack each other?

Should we use the simple product rule or the generalized one? We should use the latter! Why? Suppose we were first placing the black rook and then placing the white rook, then the possible choices for the white rook depend on where the black rook is places. As soon as you observe this kind of dependent choice in a problem, you know you have to go for the generalized product rule! There are 64 choices for the black rook. Once the black rook is placed, we cannot place the white rook in the same column or the same row as the black rook. That leaves us with 49 positions (why?), and so the answer is 64×49.

6 Being careful while counting

When you attempt counting problems, especially when you are new to them, there is a chance that you might end up *overcounting*, i.e. you may count the same element in the set more than once, hence ending up with a larger number than the actual cardinality of the set. For example, consider the following problem: how many strings binary strings of length 8 are there that contain exactly 2 zeros? Here is a faulty way that ends up overcounting: there are 8 positions to choose from for placing the first zero, and after placing the first zero, there are 7 options for the placing the second zero (the rest of the string is just filled with ones), and thus the total number is 56.
What went wrong here? We ended up counting many of the strings twice! For example, consider 10111101. There are two ways to arrive at this string following the above procedure: you could place the first zero in position 2 and the second zero in position 7, or the other way around. This makes us count this string twice!

It turns out that in the above example, the proposed method ends up counting every string (with exactly two zeros) exactly twice, and thus we can arrive at the correct number by dividing the count obtained above by two (convince yourself of this!). My suggestion to you is to double or even triple check that you are not overcounting in any of the steps of your proof, and if you are overcounting, you are scaling down the count at some later point in the proof.

Lecture 3: Division rule, subset rule and binomial coefficients, permutations with repetitions and multinomial coefficients

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 15 of the Math for CS book.

1 Recap of the two product rules

Some things to keep in mind:

- If you are counting the number of possible sequences where the choices available at the different positions in the sequence are all independent, then one can try to use the product rule. e.g., what is $|\{0,1,2\}^n|$?
- If the choices are dependent, then you cannot use the simple product rule, and might have to do a nuanced analysis. (You can't always use the generalized product rule in this case!)
- One of the scenarios where the choices are dependent in which you CAN use the generalized product rule is when you want to find the number of ways of arranging n DISTINCT objects. The answer (using the generalized product rule) is $n!$.
- In general, if you have a scenario of dependent choices where:
  - there are P_1 choices for the first entry in the sequence,
  - P_2 choices for second entry *for every choice of the first entry*,
  - P_3 choices for second entry *for every choice of the first and second entry*,
  - .
  - .
  - .
  - P_n choices for the n^{th} entry *for every choice of the first $n-1$ entries.*

  Then the total number of sequences is $P_1 \cdot P_2 \cdot \ldots \cdot P_n$.
- If the n objects are not all distinct then you CANNOT use the generalized product rule.

Question . There are 10 walls in a row. You have 15 buckets of paint, each of a different color, and a single bucket is good for coloring exactly one wall. If you want every wall to be monochromatic (i.e., a single wall should be painted using exactly one color), then how many ways are there to color the walls?

Proof. There are 15 choices for the color of the first wall, 14 choices for the color of the second wall *for every choice of the color of the first wall*, 13 choices for the third wall for every choice of the colors of the first two walls, ..., and so on, and so using the generalized product rule, we have that the number of ways to color the walls is $15 \times 14 \times 13 \ldots \times 6$. Observe that this is equal to $15!/5!$. □

In general, if you have n distinct objects in total, and want to choose and arrange r objects out of the n objects, then the number of ways of doing so is

$$n \times (n-1) \times (n-2) \times \ldots (n-(r-1)) = \frac{n!}{(n-r)!}.$$

The proof is basically a generalization of the solution to the above question: there are n choices for the first position, $n-1$ for the second for every choice of the first, and so on. We will see yet another proof in a later section.

2 Division rule/Generalized bijection method

Definition 1. *A function $f : A \to B$ is called a k-to-1 function if for every $b \in B$ there exist k distinct elements $a_1, \ldots, a_k$ such that all of them are mapped to b by f, i.e. $f(a_1) = f(a_2) = \ldots = f(a_k) = b$.*

If a function is k-to-1 with $k = 1$, then it is a bijection!

Consider the function $f : \mathbb{Z} \setminus 0 \to \mathbb{N} \setminus \{0\}$, i.e. the domain is all integers except 0 and the codomain is all natural numbers except 0, such that $f(x) = abs(x)$ where $abs(x)$ denotes the absolute value of x. Then f is a 2-to-1 function: for every non-zero natural number n, both n and $-n$ are mapped to n by f!

Lemma 2. *If A and B are finite sets, and $f : A \to B$ is a k-to-1 map, then $|A| = k \cdot |B|$.*

Proof. Let us assume that $|B| = m$ and $B = \{b_1, \ldots, b_m\}$. Note that by the definition of a function, every element in A must be map to exactly one element in B. This means we can partition the elements of A into m parts $A_1, A_2, \ldots, A_m$ where the part A_i is the set of all elements in A that get mapped to b_i by f. By the partition rule, we have that $|A| = \sum_{i=1}^{n} |A_i|$. What are the sizes of the A_is? We know that exactly k elements get mapped to every element in B, and this means that for every i, $|A_i| = k$. This means that $|A| = k \cdot m = k \cdot |B|$. This finishes the proof. □

This suggests a new method for finding the size of set B that might be hard to count directly. Find an easy to count set A such that there is a k-to-1 function from A to B for some positive integer k, and then using the above lemma, we have that $|B| = |A|/k$. This method is called the division rule or the generalized bijection method.

Question . Suppose five knights are to be seated around a round table. How many distint ways of seating them are there? Two arrangements are considered to be the same if every knight has the same two knights seated to next to them in both the arrangements.

Proof. Let B be the set of ways of seating the 5 knights around the table. Let A be the number of ways of making the 5 knights stand in a line, one behind the other. We want to find $|B|$, but instead we will find $|A|$, set up a k-to-1 function between A and B, and then use the division rule to find $|B|$.

First note that $|A| = 5!$ since it's just the number of ways of permuting 5 distinct objects. There is a natural way of converting an arrangement of the 5 knights in a line to a seating plan around the table: if the knights are standing in the order $(k_1, k_2, k_3, k_4, k_5)$, then we seat knight k_1 next to knights k_2 and k_5, knight k_2 next to knights k_3 and k_1, knight k_3 next to k_4 and k_2, knight k_4 next to k_5 and k_4, and knight k_5 next to $k1$ and $k4$. Let's call this mapping from permutations to seating arrangements $f : A \to B$.
Notice that all of the following permutations will get mapped to exactly the same seating arrangement: $(k_1, k_2, k_3, k_4, k_5)$, $(k_2, k_3, k_4, k_5, k_1)$, $(k_3, k_4, k_5, k_1, k_2)$, $(k_4, k_5, k_1, k_2, k_3)$, and $(k_5, k_1, k_2, k_3, k_4)$. In general, you can convince yourself that for every seating arrangement in B, there will be exactly 5 permutations in A that will map into it. This means that f is a 5-to-1 map and thus, using the division rule, we have that

$$|B| = \frac{|A|}{k} = 4!.$$

□

3 Subset rule and counting subsets of a set

The basic question we want to answer is as follows: suppose I have the set $\{1, 2, \ldots, n\}$, then how many ways are there to choose k numbers from this set, or, in other words, how many subsets of the given set are there that have exactly k elements? Note that it doesn't matter what order we choose the k objects in. We can assume that we are selecting all k at once so there is no inherent order in the process. Before we state the general formula, let's warm up with an example.

Question . Let $S = \{1, 2, 3, 4, 5\}$. How many subsets of S are there that have exactly 3 elements. In other words, in how many ways can you choose 3 numbers from S?

Proof. Let B be the set of subsets of S that have exactly 3 elements. Let A be the set of permutations of the elements in S. As in the last question, we will first find $|A|$, set up a function k-to-1 function f from A to B, and then find $|B|$ using the division rule.

Once again, $|A|$ is easy to find, and it is exactly 5!. Let us now give a map/function $f : A \to B$ that converts a permutation of the numbers in S to a subset of S of size 3. Here is what the map does:

- Let's say we have a permutation $(a_1, a_2, a_3, a_4, a_5)$.
- f first chops off the last 2 elements from the permutation, so we are left with (a_1, a_2, a_3).
- f then converts the remaining sequence of three elements into a set of size 3 by getting rid of the order: (a_1, a_2, a_3) becomes $\{a_1, a_2, a_3\}$.

For example, $f((1, 2, 4, 5, 3)) = \{1, 2, 4\}$. Notice that lots of different permutations will end up mapping to the same set. For example, we also have that $f(\{4, 1, 2, 3, 5\}) = \{1, 2, 4\}$, same as what we got before. The question is, how many permutations get mapped to a fixed size 3 subset in B? Let's try to compute the number of permutations that get mapped to $\{1, 2, 4\}$. Any permutation that is mapped by f to $\{1, 2, 4\}$ must have some arrangement of the numbers $1, 2, 4$ in its first three positions, and the last two positions have some arrangement of the numbers 3 and 5. Thus, the number of possible options for the first three positions are $3!$, and those for the last positions are $2!$, and so $3! \times 2!$ permutations get mapped to $\{1, 2, 4\}$ by f.

Note that in the argument above, there was nothing special about $\{1, 2, 4\}$ that we used, and the same argument would work for any size three subset. Thus, for every set in B, there are $2!3!$ permutations mapping to it, and thus f is a k-to-1 map with $k = 3!2!$. This means that

$$|B| = \frac{|A|}{k} = \frac{5!}{2!3!}.$$

□

We can use the exact same idea for deriving a general formula for the number of size k subsets of $S = \{1, \ldots, n\}$:

As before, let B be the set of all subsets of S of size k, let A be all permutations of the elements of S, and let $f : A \to B$ be a function that converts a permutation into a subset of size k in the following manner:

- Let's say we have a permutation $(a_1, \ldots, a_n)$ of S.
- f first chops off the last $n-k$ elements from the permutation, so we are left with $(a_1, \ldots, a_k)$.
- f then converts the remaining sequence of k elements into a set of size k by getting rid of the order: $(a_1, \ldots, a_k)$ becomes $\{a_1, \ldots, a_k\}$.

Now, using the same ideas as in the above proof, you can convince yourself that there are exactly $k!(n-k)!$ permutations of S that get mapped to any given size k subset of S, and thus f is a k-to-1 map. Using the division rule, and the fact that $|A| = n!$, we get that

$$|B| = \frac{n!}{k!(n-k)!}.$$

Note that there is nothing special about the objects being considered being numbers: the entire argument will go through if we were choosing, say, k students from a group of n students, or k candy bars from a set of n distinct candy bars (say each of a different brand), etc. In general:

Lemma 3. *The number of ways of choosing k objects (not giving any regard to the order) from n distinct objects is $\frac{n!}{k!(n-k)!}$.*

Definition 4. *$\binom{n}{k}$ (pronounced as "n choose k") is a* ***binomial coefficient*** *and represents the number of ways of choosing k objects from n distinct elements (without regard for order), and is equal in value to $\frac{n!}{k!(n-k)!}$.*

$\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}$ are all binomial coefficients (why didn't we also mention $\binom{n}{n+1}$?) It's not hard to see why the following is true about them (left as an exercise):

Claim 5. $\binom{n}{k} = \binom{n}{n-k}$

Question . How many binary strings of length n are there that contain exactly k ones?

Proof. To specify a binary string of length n with exactly k zeros, all I need to tell you is what positions are the k ones present at. So the number of binary strings of length n with k ones is exactly equal to the number of ways of choosing k positions from n positions, which is $\binom{n}{k}$. □

Question . How many ways are there to draw 5 cards from a deck of 52 cards so that there are at least 3 aces in the selection?

Proof. Here is a wrong way to do it. My selection must definitely have 3 aces. Let me first pick three aces that will definitely be part of my selection. The number of ways of doing that are $\binom{4}{3}$. Once I have picked three aces, I can now freely pick whatever cards I want from the remaining 49 cards, and so the total is $\binom{4}{3}\binom{49}{2}$. Why is this wrong?

To do this correctly, it's better to use the partition method: consider two cases, one when there are exactly 3 aces in the selection, and the other when all four aces are chosen. For the first cases, there are $\binom{4}{3}$ ways of choosing 3 aces from all four aces, and $\binom{48}{2}$ ways of choosing the remaining cards, and so the total is $\binom{4}{3}\binom{48}{2}$. For the second case, there is only one way to choose all four aces, are there are $\binom{48}{1} = 48$ ways of choosing the one card from the non-ace cards. The the final answer is $48 + \binom{4}{3}\binom{48}{2}$. □

3.1 Choose and permute

Recall we computed the number of ways of choosing and arranging r objects out of n distinct objects. We found it to be $\frac{n!}{(n-r)!}$, and we derived it using the generalized product rule. Here is another way to understand how to get to that number. If you have n distinct objects, and want to arrange r objects chosen from these n objects, you can think of it in two steps: first choose r objects out of n, and there are $\binom{n}{r}$ ways of doing so, and having selected the r objects you can now arrange them, and there are $r!$ ways of doing that. So, the total number is

$$\binom{n}{r} \cdot r! = \frac{n!}{r!(n-r)!} \cdot r! = \frac{n!}{(n-r)!}.$$

4 Permutations with repetitions

Question . How many distinct permutations of the word SYSTEMS are there?

Proof. Notice that the answer isn't simply 7!. The number of ways of permuting n *distinct* objects is $n!$, but we have to be more careful when we have repetitions. In this case, we have a repeat: there are three S!. We can use the division rule to solve this problem. Let B be the set of distinct permutations of the $SYSTEMS$. Let A be the set of permutations of the string $S_1YS_2TEMS_3$ (here the three Ss are distinguishable from each other). We know that $|A| = 7!$ because we are simply arranging 7 distinct objects in all possible ways.

Let us now give a map $f : A \to B$ that is k-to-1 for some value of k. We can then find $|B|$ using the division rule (so f maps a permutation of $S_1YS_2TEMS_3$ to an permutation of the letters of $SYSTEMS$). Here is how f is defined: given a permutation of $S_1YS_2TEMS_3$, f simply drops the subscripts of the three Ss in the permutation to get a permutation of $SYSTEMS$. e.g., $S_1TMS_3EYS_2$ is mapped to $STMSEYS$ by f. It's easy to see that exactly 3! different permutations of $S_1YS_2TEMS_3$ in A will map to any given permutation of $SYSTEMS$ in B — the order in which S_1, S_2, S_3 occur in a permutation does not make a difference! For example, all six of the following will map to $STMSEYS$:

- $S_1TMS_3EYS_2$
- $S_1TMS_2EYS_3$
- $S_2TMS_3EYS_1$
- $S_2TMS_1EYS_3$
- $S_3TMS_1EYS_2$
- $S_3TMS_2EYS_1$

Thus, f is a k-to-1 map with $k = 3!$, and so

$$|B| = \frac{|A|}{k} = \frac{7!}{3!}.$$

□

One can generalize this idea to derive an expression for counting permutations when there are multiple repetitions (we will omit the proof here, but the proof is a pretty straightforward extension of what we did above. Also, we will see a different and a cleaner way to derive the same expression.)

Lemma 6. *If there are n objects in total such that there are r_1 objects of type 1, r_2 objects of type 2, ..., r_k objects of type k, and $r_1 + r_2 + \ldots + r_k = n$, then the number of distinct permutations of these n objects is given by*

$$\binom{n}{r_1, r_2, \ldots, r_k} = \frac{n!}{r_1!r_2!\ldots r_k!}.$$

$\binom{n}{r_1,r_2,\ldots,r_k}$ *is called a* ***multinomial coefficient****.*

Notice that binomial coefficients are just a special case of multinomial coefficients — they are exactly the multinomial coefficients in the case when $k = 2$ and there are r objects of type 1 and $n - r$ objects of 2.

Using this, we can now compute the number of distinct permutations of words such as $ARRANGE$. There are 2 As, 2 Rs, and one copy each of N, G, and E, and so the total number is $\frac{7!}{2!2!}$.

Lecture 4: Sequences with repetitions, distributing identical objects among distinct parties, the binomial theorem, and some properties of binomial coefficients

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 15 of the Math for CS book.

1 Sequences with repetitions revisted

Let us consider the following problem to warm up towards a more general statement.

Question . How many sequences are there of the form (A, B, C), where A, B and C are disjoint subsets of $S = \{1, \ldots, 1000\}$, and $|A| = 500$, $|B| = 325, |C| = 175$?

Proof. Note that $|A| + |B| + |C| = |S|$, and thus they form a partition of S. To distribute the 1000 elements of S between A, B, and C, it suffices to specify which elements of S are chosen to go into A, and which of them are chosen to go to B — the rest will then go to C.
Let's first pick which elements will go into A. We have to choose 500 elements from S, so there are $\binom{1000}{500}$ ways of doing that. Having picked elements for A, we would be left with $1000 - 500 = 500$ elements in S, and have to choose 325 out of them for B. For this, there are $\binom{500}{325}$ ways, and so the total is:

$$\binom{1000}{500} \cdot \binom{500}{325} = \frac{1000!}{500!500!} \cdot \frac{500!}{325!175!} = \frac{1000!}{500!325!175!}.$$

□

Let us try to generalize this to the following: how many sequences are there of the form $(A_1, \ldots, A_k)$ where $A_1, \ldots, A_k$ are disjoint subsets of $S = \{1, \ldots, n\}$ and $|A_1| = r_1, |A_2| = r_2, \ldots, |A_k| = r_k$ with $r_1 + r_2 + \ldots + r_k = n$?

Let us do exactly as we did above: first pick r_1 elements to put into A_1, then r_2 to put into A_2, $\ldots$, and finally r_{k-1} elements in A_{k-1}. As soon as have filled up the first $k - 1$ elements, we will be left with $n - (r_1 + \ldots + r_k) = r_k$ elements, and thus all of them must go into A_k. Following the above calculations, we see that the number of ways of picking elements for $A_1, \ldots, A_{k-1}$ is

$$\binom{n}{r_1} \cdot \binom{n - r_1}{r_2} \cdot \binom{n - r_1 - r_2}{r_3} \cdot \ldots \cdot \binom{n - r_1 - \ldots - r_{k-2}}{r_{k-1}}$$

This can be written as

$$\frac{n!}{r_1!(n - r_1)!} \frac{(n - r_1)!}{r_2!(n - r_1 - r_2)!} \frac{(n - r_1 - r_2)!}{r_3!(n - r_1 - r_2 - r_3)!} \cdots \frac{(n - r_1 - \ldots - r_{k-2})!}{r_{k-1}!(n - r_1 - \ldots - r_{k-1})!}.$$

If you stare at this expression for a while you realize that all the numerators except $n!$ will get cancelled by some of the denominators, and the only denominators left will be $r_1!r_2!\ldots r_k!$ (where did the $r_k!$ even come from?). Thus, the expression becomes: $\frac{n!}{r_1!r_2!\ldots r_k!}$. Thus, we have

Lemma 1. *The number of sequences of the form $(A_1, \ldots, A_k)$ where $A_1, \ldots, A_k$ are disjoint subsets of $S = \{1, \ldots, n\}$ and $|A_1| = r_1, |A_2| = r_2, \ldots, |A_k| = r_k$ with $r_1 + r_2 + \ldots + r_k = n$ is $\frac{n!}{r_1!r_2!\ldots r_k!}$. In other words, the number of ways of distributing n distinct objects among k distinct parties such that party i gets exactly r_i many objects and $r_1 + \ldots + r_k = n$ is $\frac{n!}{r_1!\ldots r_k!}$.*

This also turns out to be the number of permutations when there are some repeated objects as we shall see now (the formula does look like the one we saw in the last lecture, doesn't it?)

Suppose we have n objects in total, r_1 of type 1, r_2 of type 2, ..., r_k of type k, such that $r_1+\ldots+r_k = n$. How many distinct permutations are there of these n objects?
If they were all distinct, then the answer would have been $n!$, but as we have seen before, this will lead to an overcount when we have repetitions. In the last lecture, we suggested that one can use the division rule to get around this problem, and arrive at the number of distinct permutations being

$$\binom{n}{r_1, \ldots, r_k} = \frac{n!}{r_1! \ldots r_k!},$$

where the expression in the LHS is known as a multinomial coefficient.We will now give an alternate (and much cleaner) proof of this based on what we just did above with sequence of sets.

Let's think of coming up with one possible arrangement of these n objects. What needs to be specified in order to completely specify an arrangement? Well, I need to first tell you what positions the objects of type 1 go into, then I need to tell you what positions the objects of type 2 go into, ..., and finally what positions objects of type $k - 1$ go into. As soon as I tell you all that, you'll automatically know where the objects of type k go into – they are forced to go to the remaining positions.
Now, how many ways are there to choose the positions for the objects of type 1. There are r_1 objects of type 1, so we need to choose r_1 out of n possible positions, and thus the number of ways is $\binom{n}{r_1}$. Having chosen the positions for the first type, we then choose positions for the objects of the second type. Now we need to choose r_2 positions from $n - r_1$ available ones, so that's $\binom{n-r_1}{r_2}$ ways...but, hold on? Isn't this turning out to be *exactly* what we did above?

Turns out you can do a simple trick (think bijection method): if you let A_i to be all the positions in the permutation where the r_i objects of type i will go into, then the problem becomes one of counting the number of sequences of the form $(A_1, \ldots, A_k)$, with $A_1, \ldots, A_k$ being disjoint subsets of $\{1, \ldots, n\}$ (these are the n possible positions in the permutation) of sizes $r_1, \ldots, r_k$ respectively, and $\sum_{i=1}^{k} r_i = n$. Now we can use Lemma 1 to finish the job, and conclude that the number of distinct permutations is $\frac{n!}{r_1!\ldots r_k!}$.

2 Distributing identical objects among distinct parties

Many counting problems can be rephrased as the following problem:

Question . Suppose that there are n identical candies that need to be distributed among r children (obviously, no two children are identical, and so we have r distinct parties here). In how many ways can you do this if you are not allowed to break candies into smaller pieces?

To solve these problems, it's convenient to convert them into a linear equation with constraints (think bijection method) to be able to think clearly:

- Let $x_1, \ldots, x_r$ be r variables such that x_i denotes the number of candies child i receives.
- Since all n candies must get distributed, we have that $x_1 + x_2 + \ldots + x_r = n$.
- Furthermore, since you can't break the candies and must give whole candies to the children, we must enforce that the values taken by $x_1, \ldots, x_r$ be integer, i.e. we are interested in integer valued solutions to the above equation.
- Also, it doesn't make sense if a variable becomes negative (what would that even mean? instead of giving candies to the child, ask the child for candies?), so we will say that $x_1, \ldots, x_r \geq 0$.

So the number of ways of distributing candies among children become the same as the number of integer solutions to the above equation with the additional constraint that $x_1, \ldots, x_r \geq 0$.

While this way of phrasing the problem as finding solutions to a linear equation help us think clearly and make sure we got the problem right, it doesn't really do much in terms of helping us find the answer. Instead, we will not turn to our old friends: binary strings!

Let's work with $n = 10$ and $r = 4$, i.e. 10 candies among 4 children, or equivalently, solutions to $x_1 + x_2 + x_3 + x_4 = 10$ subject to $x_1, x_2, x_3, x_4 \geq 0$. Let's look at the binary string of length 10 consisting of all zeros, and think of those zeros as candies. We can now think of putting ones into the string as the process of dividing the zeros (candies) into two parts. Say we start out with 0000000000, and add a one somewhere in there: 00010000000. This can be interpreted as there being two parts, one to the left of the one and the other to the right of the one, and the first part has 3 zeros (candies), and the second part has 7 zeros (candies). If throw in another one, then we get three parts in the same manner: 000100010000 can be thought of as three parts with part one and two having three zeros each, and the third part having 4 zeros. Of course there could be parts that are empty, e.g., in 000110000000, the second part (between the two ones) is empty.
Now, if we throw in three ones in total, we get a distribution of the 10 zeros into 4 parts. Let's fix the covention that all the zeros (candies) to the left of the first one go to x_1 (or the first child), all the zeros between the 1st and 2nd one go to x_2 (the second child), and so on. Thus, the number of ways of distributing the 10 candies between 4 children, or the number of solutions to the above equations, is exactly the number of binary strings of length 13 that have 10 zeros and 3 ones, which we know using the subset rule is $\binom{13}{3}$.

We can generalize the above argument to conclude that the number of ways of distributing n identical objects among k distinct parties is equal to the number of binary strings that have exactly n zeros and $k - 1$ ones (why $k - 1$, and not k? Because a single one created two partitions, two ones creatd three, and so on, so $k - 1$ ones will give us k partitions) which is $\binom{n+k-1}{k-1}$.

Question . In how many ways can you distribute 20 indivisible gold bars among 5 pirates so that every pirate gets at least one bar?

Proof. If we convert this an equation with constraints we get $x_1 + x_2 + \ldots + x_5 = 20$ with the constraints that $x_1 \geq 1, x_2 \geq 1, \ldots, x_5 \geq 1$. Unfortunately, the above method only works when we have constraints of the form $x_1 \geq 0, \ldots, x_5 \geq 0$. To get around this, let's first simply satisfy the bare minimum requirements of all the pirates (variables), i.e. let's given one bar to each of the pirates (variables) so that they all have at least one. This means we are left with $20 - 5 = 15$ bars.

We can now focus on distributing the remaining 15 bars among the 5 pirates without havin to worry about any additional constraints. Notice that now the equation would be simply $x_1 + \ldots + x_5 = 15$ with $x_1, \ldots, x_5 \geq 0$, and we know that the number of solutions to this equation is $\binom{19}{15}$. □

Can you generalize this?

Question . In how many ways can you distribute n indivisible gold bars among k pirates so that every pirate gets at least r bars?

Sometimes the distribution problems disguise themselves in a very inconspicuous manner (the only way to call their bluff is to practice a lot of problems). Here is an example:

Question . Suppose we roll 10 identical dice together. How many possible outcomes are there?

Proof. Note that i) all the dice are identical ii) there is no order in which we rolled the dice (all of them were rolled simultaneously). Thus, the only thing that *distinguishes* an outcome from another, or the only thing that really *defines* an outcome is the following set of statistics: how many of the dice turned up 1, how many turned up 2, ..., how many turned up 6?

Let x_i be the number of dice that turned up the number i (here $1 \leq i \leq 6$). Then we have $x_1 + \ldots + x_6 = 10$, and $x_1, \ldots, x_6 \geq 0$, and want to count the number of solutions to this set of equations. We are now on home turf, and can conclude that the answer must be $\binom{15}{10}$. □

Note that had the dice been colored with distinct colors, or if they had been rolled one at a time, then the problem would be very different, and the answer would simply be 6^{10} using the product rule (convince yourself of this!)

Here is another one (this one is from your book) that can seem totally unrelated to distribution problems:

Question S. uppose there are 20 books arranged in a row on a book shelf. In how many ways can you choose 6 books so that no two of the chosen books are adjacent to each other on the rack?

At first glance, this problem might not seem to have anything to do with what we have been discussing so far. A little bit of rephrasing makes this illusion disappear.
Note that the most obvious way of specifying a selection/choice of 6 books is to literally specify the "index" of the book if one were to start counting from 1 starting at the left most book on the shelf. A more non-obvious way would be the following: specify how many *unchosen* books there are to the left of the first chosen book, then specify how many unchosen books there are between the first and second chosen book, then the number of unchosen books between the second and third chosen book, ..., and finally the number of unchosen books to the right of the last (sixth) chosen

book. So, I can set this up in the following way: let x_1 denote the number of books to the left of the first chosen book, x_2 denote the number of books between the first and second chosen books, and so on, with x_7 representing the number of books to the right of the sixth chosen book.

What do conditions do we want to impose on these variables? Well, first of all there must be 14 unchosen books in all, and each of those 14 books must be either between two of the chosen books, or the left or right of all the chosen books, and thus: $x_1 + x_2 + \ldots + x_7 = 14$.
Furthermore, we want that no two books be adjacent, so there must be at least one unchosen book between the first and second chosen book, so we want $x_2 \geq 1$, and similarly we want $x_3, \ldots, x_6 \geq 1$. How about x_1 and x_7, there is no constraint on them: there could be as low as zero (it is possible to choose the first book from the left on the shelf, or the last book on the shelf), and so $x_1, x_7 \geq 0$. Again, we are back in familiar territory, so we can use the machinery we developed earlier in the section to finish the problem.

3 Binomial theorem

Recall that $(x+y)^2 = x^2+y^2+2xy$. Actually, there is a lot going on when one tries to derive that:

- $(x + y)^2$ is nothing but $(x + y) \times (x + y)$. Since multiplication is *distributive* over addition, there will be four terms coming from the four pssible multiplications, and then all we be added up.
- Basically, each of the four terms will consist of exactly one entry each from the two copies of $(x + y)$. For the first copy, we have the choice of picking up either x or y, and we have the same choices for the second copy. This gives $2 \times 2 = 4$ terms in total.
- The four terms are $xx + xy + yx + yy$. Now recall that $xx = x^2$ and $yy = y^2$. Also, because of multiplication being *commutative* we have that $xy = yx$, and so we get $x^2 + y^2 + 2xy$.

One can now do the same for higher powers of $(x + y)$. How about $(x + y)^n$? This is nothing but n copies of $(x + y)$ multiplied together. Again, using distributivity of multiplication over addition, we will get 2^n terms: there are n copies of $(x + y)$, and we have two choices for the first copy, two for the second copy, and so on.
Obviously, as before, many terms will "collapse" to the same expression because of commutativity ($xy = yx$, and then combining consecutive xs and ys into powers). This means that, in the final simplified expression for $(x + y)^n$, different terms will have different coefficients in front of them depending on how many of the 2^n original terms collapse to them.

Let's ask the question, what's the coefficient of $x^k y^{n-k}$? In other words, how many of the 2^n initial terms collapse to $x^k y^{n-k}$. Clearly, only those terms that have exactly k xs in them will collapse to $x^k y^{n-k}$, so how many terms have exactly k xs?
Recall that we had to pick between x and y in each of the n copies of $(x+y)$. Thus, the terms that give us $x^k y^{n-k}$ must correspond to outcomes of the picking process where we decided to pick x in exactly k of the copies of $(x + y)$, and picked y in the rest. How many such outcomes are there? It's exactly the same as deciding in which k copies of $(x + y)$ (out of the n possible) will we pick x, and there are $\binom{n}{k}$ ways of doing that. So the coefficient of $x^k y^{n-k}$ must be $\binom{n}{k}$. Since there was nothing special about n and k in the above argument, we can conclude:

Theorem 2 (Binomial theorem)**.** *Let $n \geq 0$ be an integer, then we have that*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

4 Some properties of binomial coefficients

The binomial theorem lets us derive some interesting things about binomial coefficients. Let us first ask the following question what is the sum of all binomial coefficients for a fixed n, i.e.

$$\binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n-1} + \binom{n}{n}.$$

We can answer this problem in two ways. The first one is combinatorial: recall that $\binom{n}{k}$ just denoted the number of size k subsets of the set $\{1, \ldots, n\}$. So, it seems like we are adding the number of subsets of size 0, the number of subsets of size 1, and so on up till the number of subsets of size n. But if you think about it this accounts for all possible subsets of $\{1, \ldots, n\}$, and thus the count must add up to the total number of different possible subsets of $\{1, \ldots, n\}$, which we know is 2^n.

A more algebraic way of seeing this is take the statement of the binomial theorem and set $x = y = 1$. What happens if we set $y = 1$ and $x = -1$? The LHS becomes zero, and we get

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

Now notice that $(-1)^k$ becomes 1 whenever k is even, and becomes -1 whenever k is odd, so we can write:

$$0 = \sum_{k \text{ is even and } \leq n} \binom{n}{k} - \sum_{k \text{ is odd and } \leq n} \binom{n}{k},$$

and so

$$\sum_{k \text{ is even and } \leq n} \binom{n}{k} = \sum_{k \text{ is odd and } \leq n} \binom{n}{k}.$$

This is another way of deriving something we proved using the bijection method before: $\sum_{k \text{ is even and } \leq n} \binom{n}{k}$ is nothing but the number of even size subsets of $\{1, \ldots, n\}$ and $\sum_{k \text{ is odd and } \leq n} \binom{n}{k}$ is the number of odd size subsets of $\{1, \ldots, n\}$, and the above expression tells us that there are as many odd size subsets as there are even size subsets (we gave a bijective proof for the case when n was odd, can you try for n even?).

We want to show the following:

$$\sum_{k < \frac{n}{2}} \binom{n}{k} = \sum_{k > \frac{n}{2}} \binom{n}{k},$$

i.e. the number of subsets of $\{1, \ldots, n\}$ of size less than $n/2$, is the same as the number of subsets of size more than $n/2$. I wouldn't give you the whole proof but it follows pretty simply by recalling that $\binom{n}{k} = \binom{n}{n-k}$ and so $\binom{n}{0} = \binom{n}{n}$, $\binom{n}{1} = \binom{n}{n-1}$, and so on.

Lecture 5: Multinomial theorem, some more properties of binomial coefficients, combinatorial proofs, principle of inclusion-exclusion

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 15 of the Math for CS book.

1 Multinomial theorem

The multinomial theorem is a generalization of the binomial theorem and let's us find the coefficients of terms in the expansion of $(x_1 + \dots x_k)^n$. For the sake of simplicity and clarity, let's derive the formula for the case of three variables. The more general formula is easy to guess once we have the formula for three variables.

Let's consider $(x+y+z)^{10}$. If we didn't care about combining powers and commutativity, we would get 3^{10} terms in all, terms that look like $xxxxxxxxxx$, or $xxxxyyzzyx$, etc — we have 10 copies of $(x+y+z)$ multiplied together and we must pick x, y, or z from each of the 10 copies resulting in a sequence of length 10 consisting of xs, ys, and zs.
Of course, we combine powers and use commutativity to collapse these length 10 sequences into expressions of the form $x^{k_1}y^{k_2}z^{k_3}$, where $k_1 + k_2 + k_3 = 10$. The question is what is the coefficient of $x^{k_1}y^{k_2}z^{k_3}$, or in other words, how many of the sequences "collapse" to this term? Notice that any sequence with k_1 xs, k_2 ys, and k_3 zs will collapse to this term, and so we want to understand how many such sequences are there. This is easy. There must be exactly k_1 copies of $(x+y+z)$ from which we must choose x, and there are $\binom{10}{k_1}$ ways to decide which k_1 copies to go for. From the remaining $10 - k_1$ copies we must pick y from exactly k_2 of them, and there are $\binom{10-k_1}{k_2}$ ways of doing this. Finally, we are left with $10 - k_1 - k_2 = k_3$ copies of $(x+y+z)$ and we must choose z from those (why?).
Thus the total number of sequences that collapse to $x^{k_1}y^{k_2}z^{k_3}$ is

$$\binom{10}{k_1}\binom{10-k_1}{k_2} = \frac{10!}{k_1!k_2!(10-k_1-k_2)!} = \frac{10!}{k_1!k_2!k_3!}.$$

The above formula looks very similar to the formula for counting permutations of strings that have repeated letters in them...can you see why?

More generally, let's say we had $(x+y+z)^n$, then the coefficient of $x^{k_1}y^{k_2}z^{k_3}$ for $k_1 + k_2 + k_3 = n$ in the expansion of $(x+y+z)^n$ is

$$\frac{n!}{k_1!k_2!k_3!}.$$

The above result is the **multinomial theorem**. We can consider a further generalization of the multinomial theorem where x,y, and z have a coefficient other than 1. We want to understand the coefficient of $x^{k_1}y^{k_2}z^{k_3}$ in $(ax + by + cz)^n$, where $k_1 + k_2 + k_3 = n$ and a, b, c are arbitrary real numbers. Notice that a, b, and c don't really affect the exponents of x, y, z, and so one can easily derive the formula (using the same ideas as above) for the coefficient of $x^{k_1}y^{k_2}z^{k_3}$ in this case as:

$$\frac{n!}{k_1!k_2!k_3!} \cdot a^{k_1}b^{k_2}c^{k_3}.$$

Question . Find the coefficient of $x^2y^3z^4$ in $(-3x + 2y + z)^9$.

The answer follows pretty easily from what we just discussed: $\frac{9!}{2!3!4!} \cdot (-3)^2 2^3$.

2 Yet another property of binomial coefficients

Let's fix n to be some arbitrary positive integer. We want to understand how $\binom{n}{k}$ behaves as a function of k as we vary k from 0 to n.

Lemma 1. *When $0 \leq k < \frac{n-1}{2}$,*

$$\binom{n}{k+1} > \binom{n}{k},$$

and when $\frac{n-1}{2} < k < n$,

$$\binom{n}{k+1} < \binom{n}{k}.$$

Also, $\binom{n}{k+1} = \binom{n}{k}$ only when $k = \frac{n-1}{2}$, and this can only happen when n is odd.

Proof. For $0 \leq k \leq n-1$, define r_k as follows:

$$r_k = \frac{\binom{n}{k+1}}{\binom{n}{k}},$$

i.e, the ratio of the $k+1^{th}$ binomial coefficient to the k^{th} binomial coefficient. We can simplify r_k as

$$r_k = \frac{\binom{n}{k+1}}{\binom{n}{k}} = \frac{\frac{n!}{(k+1)!(n-k-1)!}}{\frac{n!}{k!(n-k)!}} = \frac{n-k}{k+1}.$$

(Can you explain why the last equality in the above series of equalities is true?) Note that whenever $r_k > 1$, then the $k+1^{th}$ binomial coefficient is larger than the k^{th} one, and it's the other way around when $r_k < 1$, so we just need to understand the behavior of r_k with k:

$$r_k = \frac{n-k}{k+1} < 1 \Leftrightarrow k > \frac{n-1}{2},$$

and

$$r_k = \frac{n-k}{k+1} > 1 \Leftrightarrow k < \frac{n-1}{2}.$$

Also, it's easy to see that the two binomial coefficients are equal when $r_k = 1$ which can only happen when $k = (n-1)/2$. This can only happen when n is odd. This completes the proof. □

Thus, as a function of k, $\binom{n}{k}$ keeps increasing (strictly) till it becomes $\binom{n}{n/2}$ (when n is even) or $\binom{n}{(n-1)/2}$ (when n is odd). After that, in the even case, as k becomes more than $n/2$, $\binom{n}{k}$ decreases (strictly). In the odd case, $\binom{n}{(n-1)/2} = \binom{n}{(n+1)2}$, and after k goes beyond $(n+1)/2$ it strictly decreases. This means that the largest binomial coefficient in the n even case is $\binom{n}{n/2}$, and in the odd case there are two of them: $\binom{n}{(n-1)/2}$ and $\binom{n}{(n+1)/2}$.

3 Combinatorial proofs

Consider the equation:

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

We know how to show that the two are same *algebraically*, i.e. by writing out their mathematical expressions and then showing that they both reduce to the same mathematical expression.

Yet another way to show that they are equal *without* appealing to algebra is to use combinatorics:

- Let S be the number of subsets of 1 to n of size exactly k.
- We know $|S| = \binom{n}{k}$ using the subset rule that we derived in lecture.
- On the other hand, if T is the set of all subsets of 1 to n of size exactly $n-k$ we can setup a bijection between S and T: simply map a size k subset in S to its complement. By the bijection method, we know that $|S| = |T|$, and the subset rule tells us that $|T| = \binom{n}{(n-k)}$.
- But this means $\binom{n}{k} = |S| = |T| = \binom{n}{(n-k)}$.

Notice that we didn't have to use any algebra in the above proof. All we did was define S appropriately so that the expression on the left hand side of the equation we want to prove is equal to $|S|$ by using some counting method, and then we showed that using another method $|S|$ is also equal to the expression on the right hand side, and thus the two expressions must be equal. This is called a *combinatorial proof* using *double counting* (since we counted the cardinality of the same set twice using different methods).

Let us consider another example: show that $\sum_{i=0}^{n} \binom{n}{i} = 2^n$ without using an algebra, i.e. give a combinatorial proof.

Let S be the powerset (i.e. the set of all subsets) of $\{1, \ldots, n\}$. One way we can count S by using the partition method: $|S|$ = number of subsets of $\{1, \ldots, n\}$ of size 1 + number of subsets of size 2 + ... + number of subsets of size n. Since the number of subsets of $\{1, \ldots, n\}$ of size i is $\binom{n}{i}$, we get $|S| = \sum_{i=0}^{n} \binom{n}{i}$, and this show that $|S|$ is equal to the expression on the left hand side.

We now have to show that the expression of the right hand side is also equal to $|S|$. Recall that there is a bijection between the set of binary strings of length n and S: we can convert a binary string into a subset by including all those elements $1 \le i \le n$ for which the i^{th} position of x has a 1, and similarly, we can construct a binary string of length n given a subset of $\{1, \ldots, n\}$: given a subset A of $\{1, \ldots, n\}$, the string we construct for A has a 1 in the i^{th} position if i is in A otherwise there is a 0 in the i^{th} position. Since the number of binary strings of length n is 2^n, this means $|S| = 2^n$, which is the right hand side expression.

Consider another equation:

Question . Show that

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

Proof. Of course you can give an algebraic proof by putting in the formula for $\binom{n}{k}$ and then showing that the left hand side and the right hand side reduce to the same mathematical expression, but we want to give a slick combinatorial proof which will be very clean and will not involve any nasty algebra.

Let S be the set of all k size subsets of $\{1, \ldots, n\}$. Using the subset rule we know that $|S| = \binom{n}{k}$. We can partition S into two parts: all the subsets of size k that contain the element "1" (say S_1), and all the ones that don't (call it S_2). To specify a subset of $\{1, \ldots, n\}$ that contains "1", we just need to specify what other $k-1$ elements are there in the subset. There are $\binom{n-1}{k-1}$ ways of choosing those $k-1$ elements from the remaining $n-1$ elements (elements other than "1"), thus $|S_1| = \binom{n-1}{k-1}$.

Now let us count the number of subsets that don't contain a "1". In this case, we have to choose all the k elements from $n-1$ elements (all elements other than "1"), and thus the number of such sets is $\binom{n-1}{k}$, and thus that's what $|S_2|$ is equal to. Using the partition rule we know that

$$|S| = |S_1| + |S_2| = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

This means that $\binom{n}{k} = |S| = \binom{n-1}{k-1} + \binom{n-1}{k}$. □

So the general strategy is as follows: suppose you want to show that $E_1 = E_2$ where E_1, E_2 are mathematical expressions, then:

- Define a set S appropriately.
- Show that using some counting method, $|S| = E_1$.
- Show that using a different counting method, $|S| = E_2$.
- This implies that $E_1 = E_2$.

4 Principle of Inclusion-Exclusion

Suppose we have two sets $A, B \subseteq S$, and want to find $|A \cup B|$. We know that if $A \cap B = \emptyset$, then this is exactly the job for the sum rule:

$$|A \cup B| = |A| + |B|.$$

But this might not always be the case and A and B might have a nonzero intersection. What do we do then? You have seen in 205 that, more generally,

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

This takes care of the situation when we want to find the size of the union of two sets, how about three sets, i.e. find $|A \cup B \cup C|$? Let us try and derive a formula for the cardinality in this case.

Lemma 2. *Let A, B, C be subsets of S, then*

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

Proof. Let $X = B \cup C$. Then we have using the formula for union of two sets that

$$\begin{aligned} |A \cup B \cup C| &= |A \cup X| \\ &= |A| + |X| - |A \cap X| \\ &= |A| + |B \cup C| - |A \cap (B \cup C)| \\ &= |A| + |B \cup C| - |(A \cap B) \cup (A \cap C)| \qquad (1) \end{aligned}$$

We now again appeal to the formula for union of two sets to compute $|B \cup C|$ and $|(A \cap B) \cup (A \cap C)|$. We see that

$$|B \cup C| = |B| + |C| - |B \cap C|,$$

and

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

If we substitute these back into Equation 1, we get that

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

□

This is called the *principle of inclusion-exclusion*: we first overestimate $|A \cup B \cup C|$ by including too much (this is the $|A| + |B| + |C|$ part), then we underestimate it by excluding too much (this is the $-|A \cap B| - |A \cap C| - |B \cap C|$ part), and then we finally include again to get the right count for $|A \cup B \cup C|$ (this is the $|A \cap B \cap C|$ part).
In the next lecture we will see how to generalize this to the union of an arbitrary number of sets, but for now let's see some applications of the formula for union of three sets.

Question . Let $S = \{1, \ldots, 100\}$. How many numbers are there in S that are either multiples of 2 or 3 or 5?

Proof. Inclusion-exclusion comes in handy when you are dealing with a set of objects and want to know how many of those objects satisfy at least one out of two or more given conditions (in some cases, you have to count the number of objects that satisfy *all* the given conditions. We will see how to deal with those in the next lecture). In this case, the "objects" are the numbers from 1 to 100, and there are three given conditions:

1. divisible by 2
2. divisible by 3
3. divisible by 5

and we want to find the number of "objects", i.e. numbers, that satisfy at least one of the three above conditions.

The next step in solving such problems is to define a set for every condition: the set of the objects that satisfy that condition. So in this case we will define three sets for the three conditions:

1. Let A be the set of numbers in S that are divisible by 2
2. Let B be the set of numbers in S that are divisible by 3
3. Let C be the set of numbers in S that are divisible by 5

Notice that we want to find all the numbers that are either divisible by 2 or by 3 or by 5, and this translates to $|A \cup B \cup C|$ in the language of set theory ("or" is $\cup$, i.e. union, and "and" is $\cap$, i.e., intersection). Now we can apply the inclusion-exclusion formula for the union of three sets:

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

To find the cardinality of the union of the three sets, we need to find the cardinality of all the sets that occur on the right hand side of the above formula:

- $|A|$ is just the number of numbers in S that are divisble by 2, and this is just 50.
- $|B|$ is the number of numbers divisible by 3, and this is just 33.
- $|C|$ is the number of numbers divisible by 5, and this is just 20.
- $|A \cap B|$ is the number of numbers divisible by both 2 and 3 (i.e. by 6), and this is just 16.
- $|A \cap C|$ is the number of numbers divisible by both 2 and 5 (i.e. by 10), and this is just 10.
- $|B \cap C|$ is the number of numbers divisible by both 3 and 5 (i.e. by 15), and this is just 6.
- $|A \cap B \cap C|$ is the number of numbers divisible by 2, 3, and 5 (i.e. by 30), and this is just 3.

We can now substitute in all these values in the inclusion-exclusion formula we stated above, and we see that

$$|A \cup B \cup C| = 50 + 33 + 20 - 16 - 10 - 6 + 3 = 74.$$

□

In the next lecture we will see many more examples of inclusion-exclusion and will state a general formula for the union of an arbitrary number of sets.

Lecture 6: General inclusion-exclusion

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 15 of the Math for CS book.

1 General inclusion-exclusion

Last time we saw the inclusion-exclusion based formula for the cardinality of the union of three sets. What happens when there are four or more sets? We will derive the formula for union of four sets using the formula for three sets, and then I will state (without proof) the general inclusion-exclusion formula.

1.1 Inclusion-exclusion for union of four sets

Suppose we want to find $|A \cup B \cup C \cup D|$. Let's set $X = C \cup D$. Then, using the inclusion-exclusion formula for three sets we have

$$\begin{aligned}|A \cup B \cup X| &= |A| + |B| + |X| - |A \cap B| - |B \cap X| - |A \cap X| - |A \cap B \cap X| \\ &= |A| + |B| + |C \cup D| - |A \cap B| - |B \cap (C \cup D)| - |A \cap (C \cup D)| - |A \cap B \cap (C \cup D)| \\ &= |A| + |B| + |C \cup D| - |A \cap B| - |(B \cap C) \cup (B \cap D)| \\ &- |(A \cap C) \cup (A \cap D)| - |(A \cap B \cap C) \cup (A \cap B \cap D)| \qquad (1)\end{aligned}$$

The last three terms involve union of two sets and we know the formula for that. Let's compute each of them individually and put them back in Equation 1.

- $|(B \cap C) \cup (B \cap D)| = |B \cap C| + |B \cap D| - |B \cap C \cap B \cap D| = |B \cap C| + |B \cap D| - |B \cap C \cap D|$.
- Similarly,
$$|(A \cap C) \cup (A \cap D)| = |A \cap C| + |A \cap D| - |A \cap C \cap D|.$$
- $|(A \cap B \cap C) \cup (A \cap B \cap D)| = |A \cap B \cap C| + |A \cap B \cap D| - |A \cap B \cap C \cap A \cap B \cap D|$. This is the same as:
$$|(A \cap B \cap C) \cup (A \cap B \cap D)| = |A \cap B \cap C| + |A \cap B \cap D| - |A \cap B \cap C \cap D|.$$

Substituting these back into Equation 1 we get,

$$
\begin{aligned}
|A \cup B \cup C \cup D| = {} & |A| + |B| + |C| + |D| \text{ (Level I terms, } \binom{4}{1} \text{ of them)} \\
& - (|A \cap B| + |A \cap C| + |A \cap D| + |B \cap C| + |B \cap D| + |C \cap D|) \text{ (Level II terms, } \binom{4}{2} \text{ of them)} \\
& + (|A \cap B \cap C| + |A \cap B \cap D| + |B \cap C \cap D| + |A \cap C \cap D|) \text{ (Level III terms, } \binom{4}{3} \text{ of them)} \\
& - |A \cap B \cap C \cap D| \text{ (Level IV terms, } \binom{4}{4} = 1 \text{ of them).}
\end{aligned}
$$

Let's see a problem that uses inclusion-exclusion formula for four sets:

Question . Let $A = \{a_1, a_2, a_3, a_4\}$ and $B = \{b_1, b_2, b_3, b_4\}$ be two finite sets of size 4 each. A function $f : A \to B$ is said to have a fixed point if for some $i \in \{1, 2, 3, 4\}$, $f(a_i) = b_i$. How many bijective functions mapping A to B are there that have no fixed points?

Proof. As discussed in the previous lecture, we first want to identify the objects we are dealing with (numbers, subsets, functions, etc.). In this case we are asked "how many bijective functions mapping to A to B are there..", and this means the objects we are dealing with is the set of bijective functions from A to B. Let S be the set of all bijective functions from A to B. Recall that $|S| = 4!$.

Typically, in problems where we can use inclusion-exclusion, we are asked to find the number of objects (among the set of all objects we identified in first step above) that satisfy some given conditions. What are the conditions in this case? We want to find the number of bijective functions that do not have fixed points. It's always better to break down the conditions into further smaller conditions. In this case, note that not having a fixed point can be broken down into four conditions:

- $f(a_1) \neq b_1$
- $f(a_2) \neq b_2$
- $f(a_3) \neq b_3$
- $f(a_4) \neq b_4$

Once we have identified the conditions, we define a set for every condition. For $1 \leq i \leq 4$, define A_i be the set of all bijective functions for which $f(a_i) \neq b_i$. Having defined the sets, we can now translate the question being asked into the language of set theory: it is easy to see that the number of bijective functions that have no fixed points is $|A_1 \cap A_2 \cap A_3 \cap A_4|$.

Notice that we are dealing with the intersection of four sets whereas inclusion-exclusion helps us deal with the union of sets. The next step is to use the difference method to go from the intersection to union. Note that

$$|A_1 \cap \ldots \cap A_4| = |S| - |(A_1 \cap \ldots \cap A_4)^c|,$$

where $(A_1 \cap \ldots \cap A_4)^c$ is the complement of $(A_1 \cap \ldots \cap A_4)$ inside S. Using De Morgan's law, we know that $(A_1 \cap \ldots \cap A_4)^c = A_1^c \cup A_2^c \cup A_3^c \cup A_4^c$. We are now in good position: we have to compute the cardinality of the union of four sets and can use inclusion-exclusion to do so.

Notice that A_i^c (for $1 \leq i \leq 4$) is the set of all bijective functions for which $f(a_i) = b_i$. You can convince yourself that $|A_i^c| = 3!$ (that's how many ways there are to map the remaining elements in A with the remaining elements in B while making sure the function is bijective). For $i \neq j$, $A_i^c \cap A_j^c$ is the number of bijective functions for which $f(a_i) = b_i$ and $f(a_j) = b_j$. Again, it's easy to count that $|A_i^c \cap A_j^c| = 2!$. Similarly, one can show that for three distinct indices i, j, k, $|A_i^c \cap A_j^c \cap A_k^c| = 1!$, and $|A_1^c \cap A_2^c \cap A_3^c \cap A_4^c| = 1$.

We can now apply the inclusion-exclusion formula for four sets to compute $|A_1^c \cup \ldots \cup A_4^c|$. All the level I terms in the formula are equal to 4! and there are $\binom{4}{1}$ level I terms, all the level II terms are equal to 3! and there are $\binom{4}{2}$ of them, all the level III terms are equal to 1! and there are $\binom{4}{3}$ of them, and finally there is only one level IV term and it's equal to 1. If we combine all these terms with the correct signs (remember the alternation between + and − as we go from level to level), we get:

$$|A_1^c \cup \ldots \cup A_4^c| = 3! \cdot \binom{4}{1} - 2! \cdot \binom{4}{2} + 1!\binom{4}{3} - \binom{4}{4}.$$

To finish the problem, we use this to find $|A_1 \cap A_2 \cap A_3 \cap A_4|$:

$$|A_1 \cap \ldots \cap A_4| = 4! \text{ (that's } |S|) \; - \left(3! \cdot \binom{4}{1} - 2! \cdot \binom{4}{2} + 1!\binom{4}{3} - \binom{4}{4}\right).$$

□

1.2 General inclusion-exclusion formula

Suppose now we want to find $|A_1 \cup \ldots \cup A_n|$. Using induction and following the same ideas we used for the union of three and four sets, one can show that (observe the alternation of signs as we go from one level to the next!)

$$|A_1 \cup A_2 \ldots \cup A_n| = |A_1| + \ldots + |A_n| \text{ (Level 1 terms, } \binom{n}{1} \text{ of them)}$$

$$- (|A_1 \cap A_2| + \ldots + (\text{all pairs } A_i \cap A_j) + \ldots + |A_{n-1} \cap A_n|) \text{ (Level 2 terms, } \binom{n}{2} \text{ of them)}$$

$$+ (|A_1 \cap A_2 \cap A_3| + (\text{all three-wise intersections } |A_i \cap A_j \cap A_k|)\ ) \text{ (Level 3 terms, } \binom{n}{3} \text{ of them}$$

$$\ldots$$

$$(-1)^{\ell-1} ((\text{all } \ell\text{-wise intersections } |A_{i_1} \cap \ldots \cap A_{i_\ell}|)\ ) \text{ (Level } \ell \text{ terms, } \binom{n}{\ell} \text{ of them)}$$

$$\ldots$$

$$(-1)^{n-1}|A_1 \cap A_2 \ldots \cap A_n| \text{ (Level } n \text{ term, } \binom{n}{n} = 1 \text{ of them).}$$

Note that the number of terms at level ℓ is $\binom{n}{\ell}$ — it's all the ℓ-wise intersections of the n sets, and the sign at level ℓ is $(-1)^{\ell-1}$.

In the in-class exams, you will typically never need more than 4 or 5 levels of inclusion-exclusion, however your third HW has a problem that will need you to do 21 levels. The trick is to try to get a general expression for level ℓ and then sum over all the expressions for the n different levels with the appropriate sign. For example, in the problem we discussed in the previous subsection, the expression for level ℓ (for $1 \leq \ell \leq 4$) is $(4-\ell)! \cdot \binom{4}{\ell}$ (we assume that $0! = 1$), and thus we can just sum over all these expression with the right sign (the sign of level ℓ is $(-1)^{\ell-1}$). Thus, we get:

$$|A_1^c \cup A_2^c \cup A_3^c \cup A_4^c| = \sum_{\ell=1}^{4} (-1)^{\ell-1} (4-\ell)! \cdot \binom{4}{\ell}.$$

Lecture 7: Probability basics: sample space, events, probability distributions, axioms of probability, the uniform distribution

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 17 of the Math for CS book.

We will now begin the second part of the course which deals with discrete probability theory. Probability theory is the study of random/uncertain phenomenon and provides us with a framework to model these phenomenon mathematically and make predictions. Let us introduce the basic concepts of probability theory using an example:

Question . Suppose you toss a fair/unbiased coin 10 times. What is the probability that you see exactly 5 heads?

1 Experiment and sample space

Before we begin solving the problem, we need to model the random phenomenon we are dealing with. We will refer to these random phenomenon as *experiments*. The very first step is to identify what the different *outcomes* of the experiment are.

In the case of the above question, the experiment is tossing a fair coin 10 times, and so the possible outcomes are all possible sequences of length 10 where every entry in the sequence is either "H" (heads) or "T" (tails). For example, one possible outcome is $HHHHHHHHHH$ which simply denotes the outcome in which we see a heads every single time, and another example is $HTTHHHHTTT$, which says that we see a heads in the 1st, 4th, 5th, 6th, and 7th coin toss and tails in the 2nd, 3rd, 8th, 9th, and 10th coin toss.

Note the problem seems to suggest that we are interested only in the total number of heads we see during the 10 coin tosses, so we can also define an outcome to be simply the number of heads we see during the 10 trials. While this is a totally legitimate choice, it does *miss out* on other potentially useful details about how the experiment *unfolded* which might be useful or even critical to compute probabilities later on.
For example, in the 10 coin tosses case, if you define an outcome to be just the total number of heads you see during the trials, then you miss out on the information about where (i.e., during which trials) the heads came up! Thus, **it's always safer to define an outcome in a way that it captures all the important details about how the experiment unfolded**.

Anyway, the set of all possible outcomes is called the *sample space*, and is denoted using the Greek letter Ω. So in the case of the coin toss problem, we have

$$\Omega = \{H, T\}^{10},$$

where $\{H, T\}^{10}$ denotes the set of all sequences of length 10 consisting of "H" and "T". The size of the sample space, thus, is 2^{10}. Often it's useful to represent a sequence of coin tosses using binary strings, usiing 1 to represent heads and 0 to represent tails. So in this case, we could have defined the sample space to the set of all binary strings of length 10.

2 Events

An *event* is a subset of the sample space, i.e. $E \subseteq \Omega$. This means that an event is just a set containing *some of* the outcomes (Ω, the sample space, is the set of ALL possible outcomes) If an event E is a *singleton set*, i.e. it contains exactly one outcome from Ω, the sample space, then it's called an *atomic event* or an *elementary event.* For example, consider the event E that corresponds to seeing 10 heads during the 10 trials, then $E = \{HHHHHHHHHH\}$, and E is an atomic/elementary event.

Events that contain more than one outcome are called *compound events.* For example, consider the event E of seeing a heads in the first coin toss, then E is the set consisting of all sequences of length 10 consisting of H and T that begin with an H, and $|E| = 2^9$ (why?). In this case, E is a compound event.

We say an event *takes place or happens* if the experiment results/ends up in any one of the outcomes in that event.

What is the set of all possible events? It's the powerset of the sample space, i.e. 2^Ω or $powerset(\Omega)$. (Why?)

2.1 Operations on events

Since events are nothing but subsets of the sample space Ω, we can use the usual set operations on them. Let Ω be some sample space, then

- if $E \subseteq \Omega$ is an event, the complement of the event E is simply the set $E^c = \Omega - E$. For example, going back to the 10 coin tosses example, let E be the event that you see at least one heads during the 10 trials. As in the case of counting, it's much easier to deal with the complement of E, i.e. the event that you see no heads and thus all tails. Thus, in this case, $E^c = \{TTTTTTTTTT\}$, and $|E| = |\Omega| - |E^c| = 2^{10} - 1$ (think difference method!). Another way of to think of E^c is to think of it as the set of all outcomes where the event E doesn't "happen".
- if E_1 and E_2 are two events in Ω, then we can define the event $E_1 \cup E_2$, i.e. the set of all outcomes that are either in E_1 or in E_2 or in both. For example, if E_1 is the event of seeing at least one heads and E_2 is the event of seeing no heads, then $E_1 \cup E_2$ is the event of seeing zero or more heads which is basically Ω, the entire sample space.
- Similarly, one can define $E_1 \cap E_2$, the set of all outcomes that are in both E_1 and E_2. Suppose E_1 was the set of outcomes in which you see at least one heads, and E_2 was the set of outcomes where you see at least one tails, then $E_1 \cap E_2$ is the set of all outcomes where you see at least one heads AND at least one tails. (Can you compute $|E_1 \cap E_2|$?)
- Finally, we can talk about $E = \emptyset$, the empty event that doesn't contain any outcome and *never* happens (why?), and $E = \Omega$, the trivial event that always *happens* (why?).

2.2 Mutually exclusive events/disjoint events

If we have a bunch of events $E_1, \ldots, E_k$ such that for every pairs of sets in this bunch E_i, E_j we have that $E_i \cap E_j = \emptyset$, then we say that $E_1, \ldots, E_k$ are *mutually exclusive or disjoint* events. This means that if an event E_i happens/takes place, then none of the other events among the remaining $k-1$ events take place or happen. For example, if E_1 is the event of seeing at least two heads during the 10 trials, and E_2 is the event of seeing exactly one heads during the trials, and E_3 is the set of all outcomes where you don't see any heads, then E_1, E_2, E_3 are mutually exclusive events, and in fact $E_1 \cup E_2 \cup E_3 = \Omega$ (why?).

3 Probability

Let's go back to the question we are trying to solve. The question asks what is the probability of seeing exactly 5 heads. We have identified what the outcomes are and what the sample space is. We now want to define a probability distribution on our sample space. What is a probability distribution? It's a function that assigns a real number to every event of our sample space. The real number is supposed to be a "measure of likelihood" of that event.

More formally the probability distribution is a function $P : powerset(\Omega) \rightarrow \mathbb{R}$. Why is the domain of the function P $powerset(\Omega)$? It's because we want the function P is supposed to associate a real value with every event, and its domain, thus, must be the set of all events which is $powerset(\Omega)$ or 2^Ω.

What behavior do we expect the probability function to have? We want it to model our intuitive notion of "likelihood" or "chance".

3.1 Axioms of probability

Here are some common sense properties we want the probability distribution function P to satisfy:

1. Obviously, it doesn't make sense to use negative numbers to denote the likelihood of an event or the chance of the event happening (What would it even mean to say there is a -50% chance of raining tomorrow?). So the first condition we want is that for every event E,

$$P(E) \geq 0.$$

2. Remember, when we defined the event $E = \Omega$ and said it's the event that *always happens.* Thus, we want to have that

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

   Similarly, remember the event $E = \emptyset$, the event that has no outcomes and thus should never happen? It makes sense to have $P(\emptyset) = 0$. (it turns out we can *derive* this fact from the other axioms — we will see this later)

3. Suppose E_1 is the event of seeing at least two heads during the 10 trials and E_2 is the event of seeing exactly 1 heads during the trial. First note that E_1 and E_2 are disjoint. Also, if the chance of E_1 happening is p_1 and the chance of E_2 happening is p_2, we *expect* the chance of

either E_1 happening or E_2 happening to be $p_1 + p_2$. We can formalize this to the following: If $E_1, \ldots, E_k$ are disjoint events/mutually exclusive events, then

$$P(E_1 \cup \ldots \cup E_k) = P(E_1) + P(E_2) + \ldots + P(E_k).$$

We can all this axiom the *sum rule of probability.*

3.2 Uniform distribution for equally likely outcomes

Ok, so now we understand that we need to define a probability distribution on the sample space Ω which in the case of the coin toss problem is the set of all binary strings of length 10 (or the set of all sequences of length 10 made using H and T). How should out probability distribution look, or in other words what numbers should we associate with the all the events?

Let's try to first define the probability for the atomic/elementary events. There are 2^{10} atomic events, let's denote them by $E_1, \ldots, E_{2^{10}}$. Intuitively, since the coins are all unbiased and fair, we expect all the outcomes, and thus all the elementary events, to be equally likely! (convince yourself!). And so we want:

$$P(E_1) = P(E_2) = \ldots = P(E_{2^{10}}).$$

We know that $P(\Omega) = 1$, and also that $\Omega = E_1 \cup \ldots \cup E_{2^{10}}$. The last fact is true because the sample space Ω is made up of these elementary/atomic events and nothing else, so we can write Ω as a union of these events. Moreover all these elementary events are disjoint because remember: each of these elementary events contain exactly one outcome, and no two events contain the same outcome (otherwise they would be the same event!). We can thus use the sum rule of probability:

$$1 = P(\Omega) = \sum_{i=1}^{2^{10}} P(E_i).$$

Let $p \geq 0$ be such that

$$P(E_1) = P(E_2) = \ldots = P(E_{2^{10}}) = p.$$

Then, we have using the above equation that

$$1 = \sum_{i=1}^{2^{10}} P(E_i) = 2^{10} \times p \implies p = \frac{1}{2^{10}},$$

and thus,

$$P(E_1) = P(E_2) = \ldots = P(E_{2^{10}}) = p = \frac{1}{2^{10}}.$$

Thus, our probability distribution should assign $1/2^{10}$ to every elementary/atomic event!

But so far we have just assigned probabilities for atomic events! How about compound events? Let E be an arbitrary compound event of size k that consists of the outcomes $o_1, \ldots, o_k$, i.e, $E = \{o_1, \ldots, o_k\}$. Then we can write

$$E = \{o_1\} \cup \{o_2\} \cup \ldots \{o_k\},$$

where $\{o_1\}, \ldots, \{o_k\}$ are nothing but elementary/atomic events! Using the sum rule for probability

$$P(E) = \sum_{i=1}^{k} P(\{o_i\}).$$

Recall that every elementary event was given a probability of $1/2^{10}$, and so

$$P(E) = \sum_{i=1}^{k} P(\{o_i\}) = \frac{k}{2^{10}} = \frac{|E|}{2^{10}}.$$

So we have now completely specified the probability distribution P. It's not hard to verify that the P defined above satisfies the axioms of probability:

1. For every event E, since $|E| \geq 0$, we have that $P(E) = \frac{|E|}{2^{10}} \geq 0$
2. Also, $P(\Omega) = \frac{|\Omega|}{2^{10}} = \frac{2^{10}}{2^{10}} = 1$, and similarly $P(\emptyset) = \frac{|\emptyset|}{2^{10}} = 0$.
3. Let E_1 and E_2 be disjoint events of Ω, then we know using the sum rule for sets that $|E_1 \cup E_2| = |E_1| + |E_2|$. It follows that

$$P(E_1 \cup E_2) = \frac{|E_1 \cup E_2|}{2^{10}} = \frac{|E_1| + |E_2|}{2^{10}} = \frac{|E_1|}{2^{10}} + \frac{|E_2|}{2^{10}} = P(E_1) + P(E_2).$$

This sort of a probability distribution that assigns the same probability to every elementary event/outcome is called the *uniform distribution.*

In general, for any experiment and any sample space Ω associated with it, the uniform distribution on Ω is the one that assigns probability $\frac{1}{|\Omega|}$ to all the elementary/atomic events of Ω, and for a compound event E assigns the probability $\frac{|E|}{|\Omega|}$.

Whenever you are dealing with an experiment where all the outcomes should be equally likely, use the uniform distribution!

3.3 Solution to the question

We are now in perfect shape to answer the question: what is the probability of seeing exactly 5 heads during the 10 trials? Well, our Ω was defined to be the set of all binary strings of length 10, and we used the uniform probability distribution on Ω. We are interested in the probability of the event E that we see exactly 5 heads. What outcomes does E contain? E contains all the binary strings that have exactly 5 ones, and thus $|E| = \binom{10}{5}$. Thus, the uniform distribution would give us

$$P(E) = \frac{|E|}{|\Omega|} = \frac{\binom{10}{5}}{2^{10}}.$$

Here is another question to think about:

Question . If we roll two fair dice, one black and one white, what is the probability that both dice turn up 3? What is the probability that both dice turn up 3 or both dice turn up 1?

It's not hard to see that Ω is $\{1,\ldots,6\}\times\{1,\ldots,6\}$, i.e. the set of all 2-tuples where every entry is a number between 1 and 6, and the first entry in the tuple represents the number that the black dice rolls, and the second entry is the number that the white dice rolls. Thus $\Omega = 36$. Let $E_1 = (3,3)$ and $E_2 = (1,1),(3,3)$. Then E_1 is the set of outcomes where both dice roll 3, and E_2 is the event where both dice roll 3 or both dice roll 1. It follows from out discussion above that

$$P(E_1) = \frac{1}{36}$$

and

$$P(E_2) = \frac{2}{36}.$$

3.4 Some more properties of probability distributions

Let us fix an experiment and associate a fixed sample space Ω with it. Furthermore, let us say that we have a probability distribution P on Ω. Let $E \subseteq \Omega$ be an event. Suppose we know the value of $P(E)$. What is $P(E^c)$?

Recall that $E^c = \Omega - E$, and thus we have that E and E^c are disjoint events and $E \cup E^c = \Omega$. It follows from the sum rule of probability that

$$P(\Omega) = P(E \cup E^c) = P(E) + P(E^c).$$

We also know that $P(\Omega) = 1$, and so we get that

$$P(E) + P(E^c) = 1 \implies P(E^c) = 1 - P(E).$$

Thus: the probability of the event not happening is one minus the probability of the event happening.

Using this, we can also conclude the sort-of-obvious fact that for every event E, $P(E) \leq 1$. This is because we know that $P(E) + P(E^c) = 1$, and we also know that $P(E^c) \geq 0$, and so $P(E) \leq 1$. Thus, combining this with one of the axioms we saw earlier, we have that for every event E,

$$0 \leq P(E) \leq 1.$$

We know how to compute $P(E_1 \cup E_2)$ when E_1 and E_2 are disjoint, but how about the case when they are not? Turns out one can prove an inclusion-exclusion formula for probability distributions:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

To see why this is true, observe that $E_1 - E_2$, $E_2 - E_1$, and $E_1 \cap E_2$ are all disjoint sets, and we can write

$$E_1 \cup E_2 = (E_1 - E_2) \cup (E_2 - E_1) \cup (E_1 \cap E_2).$$

Thus using the sum rule of probabilities, we get

$$P(E_1 \cup E_2) = P(E_1 - E_2) + P(E_2 - E_1) + P(E_1 \cap E_2).$$

We also know that $P(E_1) = P(E_1 - E_2) + P(E_1 \cap E_2)$ (why?) and $P(E_2) = P(E_2 - E_1) + P(E_1 \cap E_2)$, and thus we can substitute $P(E_1 - E_2) = P(E_1) - P(E_1 \cap E_2)$ and $P(E_2 - E_1) = P(E_2) - P(E_1 \cap E_2)$ in the above expression for $P(E_1 \cup E_2)$. This gives us

$$P(E_1 \cup E_2) = P(E_1) - P(E_1 \cap E_2) + P(E_2) - P(E_1 \cap E_2) + P(E_1 \cap E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

Once we have the inclusion exclusion formula for the probability of union of two events, using the same ideas as in the case of cardinality of union of sets we can find a general formula for the probability of union of k events.

Using the inclusion-exclusion formula, we can derive an important ineqality:

$$P(E_1 \cup E_2) \leq P(E_1) + P(E_2).$$

This is called the *union bound.* The proof is really simple: we know that $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$, and that $P(E_1 \cap E_2) \geq 0$ (why?), and thus $P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$.

Suppose that we have events E_1, E_2 such that $E_1 \subseteq E_2$. How does $P(E_1)$ compare with $P(E_2)$? Intuitively, whenever E_1 "happens", E_2 also happens, and thus it should be the case that $P(E_1) \leq P(E_2)$. We can prove this formally: we can write $P(E_2) = P(E_1) + P(E_2 - E_1)$ (why?), and observe that $P(E_2 - E_1) \geq 0$, and this gives us that $P(E_2) \geq P(E_1)$!

Lecture 8: Conditional probability I: definition, independence, the tree method, sampling, chain rule for independent events

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 18 of the Math for CS book.

Before we begin the lecture on conditional probability, I want to mention two tips:

1. For a counting problem, unless the problem explicitly says *identical* objects, don't assume the objects are identical.

2. For a probability problem, always assume that the objects are distinct, *even if* the problem says they are identical. This is because when computing the probability we are concerned with *the number of ways of reaching an outcome* and not just the outcomes themselves. For example, if a bucket contains 10 identical white balls, and 20 identical red balls, then the probability of picking a red ball (if you pick a ball at random) is 2/3. This is because the number of ways of reaching the outcome of seeing a red ball is 20, and the total number of outcomes is 30. If you just assume that even identical objects are distinct, you don't have to worry about this, since the distinction between "ways" and "outcomes" disappears. This trick works in 99% of the cases.

1 Definition of conditional probability

Let us assume Ω is a sample space for an experiment, and A and B are events. Let us motivate conditional probability through two examples:

Question . I want to draw two cards from a complete deck of shuffled cards, one at a time. If the first card I draw is an ace, what is the probability that the second card is also an ace?

In this case, let B be the event that the first card I drew was an ace, and let A be the event that the second card I draw will be an ace. I have been given the information that the event B has happened/occurred. If I was not given this information, what would be the probability of A? We would define the sample space Ω as set of all sequences of length 2 where every entry of the sequence is one of the 52 cards and no repetition is allowed, and thus $|\Omega| = \binom{52}{2}2!$. A would be the set of sequences of two cards (without repetition) so that the second card is an ace. Thus, $|A| = \binom{4}{1}\binom{51}{1}$ (why?), and thus

$$P(A) = \frac{|A|}{|\Omega|} = \frac{4 \times 51}{52 \times 51} = \frac{4}{52},$$

since all outcomes are equally likely and we use the uniform distribution.

But now you have been given the information that B occured, i.e. the first card that was picked was an ace. Does this affect the probability of A happening? Intuitively, it does because when we are picking the second card we have one less ace to pick from. We represent this "new" probability, i.e. the probability of A happening *given* that B has happened by $P(A|B)$ ("probability of A given B").

To compute this probability, note that when picking the second card, our "effective" sample space has changed: only outcomes within $B \subseteq \Omega$ are now possible. If this is the case, then only those outcomes in A that are *also* contained in B can happen, and thus, intuitively,

$$P(A|B) = \frac{|A \cap B|}{|B|}.$$

It turns out this is indeed how conditional probability of A given B is defined in the uniform case. More generally,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

We can now use this formula to compute $P(A|B)$ for the above problem. To do this, we need to find $|A \cap B|$ and $|B|$. You can convince yourself that $|B| = 4 \times 51$. As for $|A \cap B|$, i.e. the number of sequences of cards where both the first and second card are aces, it's 4×3. Thus,

$$P(A|B) = \frac{4 \times 3}{4 \times 51} = \frac{3}{51}.$$

This also makes intuitive sense: after you have picked an ace in the first draw, there are 3 aces left (those are the outcomes we want) and the total outcomes is the total number of cards left which is 51, and thus the probability should be 3/51.

Let us consider a different scenario where conditional probabilities can be defined and used.

Question . I drew two cards from a complete deck of shuffled cards, one at a time. If the second card I drew is an ace, what is the probability that the first card was also an ace?

Here we can assume that the experiment has happened, and that the second card that was drawn was an ace, and we want to do a "retrospective analysis" and compute how likely it is that the first card was also an ace *given* the information that the second card was an ace? Let A be the event that the second card is an ace, and let B be the event that the first card is an ace. If you have no information about A happening, then you would say that the probability that B happened is just 4/52 (why? see above calculations). But now that you have been told the information that A did happen, what is the probability that B happened. Again, we denote this conditional probability by $P(B|A)$, and following the same intuition as above, we can say that

$$P(B|A) = \frac{|A \cap B|}{|A|}$$

in the uniform distribution case, and in general

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

You can use the above formulas to conclude that in this case $P(B|A) = \frac{3}{51}$. This means that with this new information (that A did happen) you "updated" your probability that B happened (we will see more about updates later on).

Thus, these are the two scenarios in which conditional probability is useful: the first one is a scenario where there is a multi-step experiment, and you told that some event that involves the previous steps has happened, and you want to "predict" how likely it is that some event that involves future steps will happen, and the second one is where the experiment has already occured and you are doing a retrospective analysis (you are told some event B happened for sure, and want to know the probability that A happened in the light of this information).

1.1 A word of caution

In the above example, it turned out that $P(A|B) = P(B|A)$ but this is not generally true because $P(A|B) = P(A \cap B)/P(B)$ and $P(B|A) = P(A \cap B)/P(A)$, and thus the two are equal if and only if $P(A) = P(B)$.

1.2 Comparing $P(A)$ with $P(A|B)$

In general, $P(A|B)$ could larger or smaller as compared to $P(A)$. Consider the case when $B \subseteq A$. In this case, if I tell that the experiment resulted in an outcome of the event B, then it also resulted in an outcome of the event B, and thus $P(A|B) = 1 \geq P(A)$. In the other extreme, let's say A and B are mutually exclusive events, i.e. if one happens then the other cannot happen. In this case, $P(A|B) = 0$ (why?), and thus $P(A|B) <= P(A)$.

What can we say if $P(A|B) = P(A)$? Intuitively, it means that B happening does not affect the probability of A happening, and thus in some sense the two events are "uncorrelated" (in the colloquial sense).

2 Independent events

If A, B are two events of a sample space Ω such that $P(A|B) = P(A)$, then they are called independent events. Note that if $P(A|B) = P(A)$ then $P(A \cap B)/P(B) = P(A)$ and so

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

It turns out we can also define two events to be independent if $P(A \cap B) = P(A)P(B)$ and the two definitions are equivalent (you can derive one assuming the other is true), and the latter is the preferred definition since in certain cases when $P(B) = 0$, $P(A|B)$ might not be defined.

Obviously, if $P(A|B) = P(A)$ it's also true that $P(B|A) = P(B)$ (again, you can derive one from the other).

Question . Suppose we roll a black and a white dice (both fair). Let A be the event that the white dice rolls one, and let B be the event that the sum of the numbers rolled by the dice is 7. Show that A and B are independent.

Proof. In this case Ω is the set of all ordered pairs (w, b) where the first component represents the number rolled by the white dice and the second one represents the number rolled by the black dice.

It is easy to see, then, that

$$B = \{(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)\}$$

and thus $|B| = 6$. Within B, the outcomes that have the white dice rolling a one is just $(1,6)$, and thus $|A \cap B| = 1$, and so

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{1}{6}.$$

It is not hard to show that $|A| = 6$ and $|\Omega| = 36$, and so

$$P(A) = \frac{1}{6} = P(A|B).$$

This completes the proof that A and B are independent. □

2.1 Mutual independence

What does independence look like when we are talking about a whole bunch of events, say $A_1, \ldots, A_n$ in a sample space Ω. We say that $A_1, \ldots, A_n$ are *mutually independent* if for every $1 \le i \le n$, the probability of A_i happening is unaffected by the occurence of a subset of the n events (obviously, the subset doesn't include A_i itself). For example, say $n = 7$ then the probability of A_3 happening is the same as the probability of A_3 happening given that A_4, A_6, A_7 have happened, etc.

Formally, we say that $A_1, \ldots, A_n$ are mutually independent if

$$\begin{gathered}
P(A_1 \cap A_2) = P(A_1)P(A_2) \\
P(A_1 \cap A_3) = P(A_1)P(A_3) \\
\vdots \\
P(A_1 \cap A_n) = P(A_1)P(A_n) \\
P(A_2 \cap A_3) = P(A_2)P(A_3) \\
\vdots \\
P(A_{n-1} \cap A_n) = P(A_{n-1})P(A_n) \\
P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) \\
\vdots \\
P(A_{i_1} \cap \ldots A_{i_k}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_k}) \\
\vdots \\
P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot \ldots \cdot P(A_n).
\end{gathered}$$

That is, if you take any subset of $A_1, \ldots, A_n$, then the probability of their intersection if the product of the individual probabilities of the events in the subset.

If someone says $A_1, \ldots, A_n$ are independent events, they mean they are mutually independent. We will see other notions of independence in future lectures.

2.2 The chain rule for independent events

Suppose Ω is a sample space, and E is an event of Ω. Furthermore, let's assume that $E = A_1 \cap A_2 \cap \ldots \cap A_n$, where $A_1, \ldots, A_n$ are mutually independent events. In many such situations, it's hard to directly compute $P(E)$, but it's easy to compute each of $P(A_1), \ldots, P(A_n)$. We can exploit the fact that $A_1, \ldots, A_n$ are independent, and say that

$$P(E) = P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2) \ldots P(A_n).$$

Consider the following problem

Question . 20 fair dice are tossed. What is the probability the product of the numbers rolled by the 20 dice is odd?

Proof. Ω in this case is the set of all sequences of 20 coin tosses or the set of all binary strings of length 20 where 1 represents heads and 0 represents tails. E is the event that the product is odd. Note that the product of the numbers rolled by the 20 dice is odd if and only if all 20 dice roll an odd number. Let $A_1, \ldots, A_{20}$ be such that A_i denotes the probability that the i^{th} dice rolls an odd number. Then we can write $E = A_1 \cap A_2 \cap \ldots \cap A_{20}$. Furthermore, $A_1, \ldots, A_n$ are all independent, and so

$$P(E) = P(A_1 \cap \ldots \cap A_{20}) = P(A_1)P(A_2) \ldots P(A_{20}).$$

Note that for every dice i we have that $P(A_i) = 1/2$ since out of the six possible outcomes exactly three lead to an odd number being rolled. This means that $P(E) = (1/2)^{100}$. □

3 The tree method for multi-stage experiments

The tree method is useful for analyzing experiments which have multiple stages: "first do this, if the outcome is ... then do that, if the outcome of that is ... then do ...". For example, consider the following question:

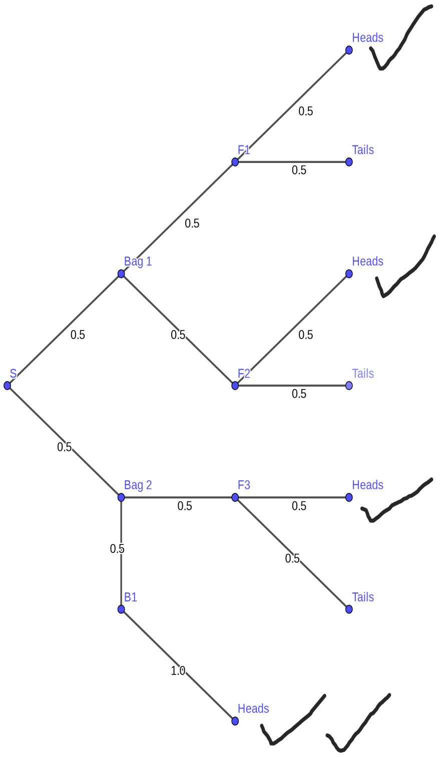
Question . I have two bags. Bag 1 contains two fair coins and Bag 2 contains a fair coin and a coin with two heads. I toss a fair coin. If the outcome is heads, I pick Bag 1, and if the outcome is tails I pick Bag 2. Then from the bag in picked I the previous step, I randomly pick a coin such that both coins in the bag are equally likely to be picked. Finally I toss the coin I picked from the bag. What is the probability that the final outcome is heads? Suppose I tell you that the final outcome of the experiment was heads how likely is it that the final coin toss was done with the coin with two heads?

Proof. Let us assume that Bag 1 contains two fair coins F_1 and F_2 and Bag 2 contains F_3 and B_1, where B_1 is the the coin with two heads. The experiment happens in three stages:

1. You first coin to decide which Bag to go for.
2. Based on what coin you got, you select a bag, and then randomly choose one of the coins in the bag.

3. Finally you toss the coin you took out of the bag.

We can represent the various stages of the experiment and their outcomes using the following tree:



Here S denotes the start of the experiment. The probability of a branch is written along the branch. For example, from S we branch into two possible outcomes with equal probabilities (1/2 each): either pick Bag 1 or pick Bag 2.

Once we have the tree drawn along with all the probabilities written on the branches, we can now identify the outcomes we are interested in. Notice that the outcomes of the final stage are the leaves of this tree. Since we are interested in the probability that we see a heads in the final stage, let's first put a check mark on all those leaves which correspond to seeing a heads. The probability of arriving at a particular leaf is the product of all the probabilities along the path from S to that leaf.
For example, consider the heads leaf on the very top of the figure. The probability of arriving at that leaf is $0.5 \times 0.5 \times 0.5 = 0.125$. Similarly, for the second leaf from the top that is labelled heads, the probability is again 0.125, and it's the same for the third leaf from the top that is labelled "heads". For the bottom-most leaf that is labelled heads, the probability is $0.5 \times 0.5 = 0.25$. Thus, the total probability of seeing a heads in the final stage of the experiment is $0.125 + 0.125 + 0.125 + 0.25 = 0.625$. If B is the event of seeing heads in the final coin toss, we have that $P(B) = 0.625$.

Let A denote the probability of using the two-headed coin (i.e., B_1) in the last stage. We want to compute $P(A|B)$. To do this, we first identify the leaves whose root-to-leaf paths pass through $B1$ (this corresponds to the outcome of using a two-headed coin) among those leaves that are labelled "heads" and put another check mark near it. In our case, there is only one such leaf, and the probability of that leaf is $0.5 \times 0.5 \times 1 = 0.25$. This is basically $P(A \cap B)$. We can now compute

$P(A|B)$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.25}{0.625} = 0.4.$$

□

In general, after putting down the second round of check marks, you would add up the probabilities of all the leafs that have double check marks and divide that by the sum of the probabilities of all the leaves with a single check mark to get the desired conditional probability.

4 Sampling

Suppose you have an urn that contains n balls in it. Each ball has a unique number between 1 and n printed on it. *Sampling* is nothing but picking balls from the urn. There are two kinds of sampling. Let us say you want to sample k balls from the urn:

1. You can take k balls at once, or
2. You can take one ball at a time.

For the latter type, we have to possibilities:

1. Once you pick a ball, you put it back into the urn so that it may be picked again later, or
2. Once you pick a ball, you don't put it back, so that every ball may be picked at most once.

If you sample one at a time, and put the ball back after it is picked, this is called *sampling with replacement*, and if you sample balls one at a time, and don't put back the balls that are already picked, then this is called *sampling without replacement.*

Sometimes it's helpful to recognize that the problem you are dealing with is basically a sampling problems. Once you figure that out, you need to figure out whether you are picking objects at once, or one-by-one. If it's the latter you should figure out whether it's with or without replacement. Once you have nailed down these details, you are in a good shape because the solutions to many sampling problems follow the same template.

Question . Every time you buy a bag of chips from the store you are equally likely to find one of 10 distinct toys in the bag. If you buy 5 bags of chips, what's the probability that at least two of the bags contain the same toy?

Proof. This is a sampling problem in disguise. Every time we buy a bag of chips we are basically sampling from the set of toys, and are equally likely to sample one of 10 distinct toys. Is this with or without replacement? It's with replacement because it is possible sample the same toy more than once.

Let's define our outcomes as sequences of length 5, where every entry can be a number between 1 and 10 (representing the 10 different toys), and the choices for every entry are independent. Thus,

$|\Omega| = (10)^5$. Let A be the set of outcomes that correspond to picking at least one toy twice. Then A^c is the set of all outcomes where we pick 5 different toys. The number of sequences of length 5 where every entry is a number between 1 and 10, and no repetition is allowed, is $\binom{10}{5}5!$ (choose 5 out of 10 and then permute them). Thus, $|A^c| = \binom{10}{5}5!$. This means

$$P(A) = 1 - P(A^c) = 1 - \frac{\binom{10}{5}5!}{(10)^5}.$$

□

Here is another question, but this one deals with *sampling without replacement*:

Question . A box has 6 identical yellow bulbs, 5 identical blue bulbs, and 4 identical white bulbs. Bulbs are sampled one-by-one without replacement. What is the probability that the third bulb that's sampled is not white?

Proof. Let's first use the tip we discussed at the beginning of this lecture: make identical objects distinct! So let's assume we have 15 distinct bulbs: $Y_1, \dots, Y_6, B_1, \dots, B_5, W_1, \dots, W_4$. Since we are sampling one at a time without replacement, a good choice for the sample space if the set of all possible permutations of $Y_1, \dots, Y_6, B_1, \dots, B_5, W_1, \dots, W_4$ – a permutation represents the order in which the bulbs are picked. Thus, $|\Omega| = 15!$[1].

We are interested in the event E that consists of all permutations in which a white bulb *does not* appear in the third position. Then E^c is the set of all permutations in which a white bulb does occur in the third position. There are 4 choices for deciding which white bulb to have in the third position, and there are 14! ways to arrange the remaining bulbs, thus $|E^c| = 4(14!)$. This implies that

$$P(E) = 1 - P(E^c) = 1 - \frac{4(14!)}{15!} = 1 - \frac{4}{15} = \frac{11}{15}.$$

□

Question . A box has 6 identical yellow bulbs, 5 identical blue bulbs, and 4 identical white bulbs. You pick 5 random bulbs from the box. What is the probability that at least one yellow bulb was picked?

Proof. Again, we make all the bulbs distinct: so we have 15 distinct bulbs now (as in the previous problem). This problem can be modeled as sampling 5 bulbs at once from the box, and so the sample space should be the set of all possible subsets of $Y_1, \dots, Y_6, B_1, \dots, B_5, W_1, \dots, W_4$ of size 5. There are $\binom{15}{5}$ such subsets, and so $|\Omega| = \binom{15}{5}$.

Let E denote the event that at least one yellow bulb is picked. Then E^c is the event that no yellow bulb is picked. There are exactly $\binom{9}{5}$ subsets that don't contain any of the yellow bulbs and so $|E^c| = \binom{9}{5}$. This means that

$$P(E) = 1 - P(E^c) = 1 - \frac{\binom{9}{5}}{\binom{15}{5}} = 1 - \frac{9 \times 8 \times \ldots \times 6}{15 \times 14 \times \ldots \times 11}.$$

[1] Some of you might have other ways of solving this problem. That's totally fine. For those who still have issues with solving probability problems, I recommend following my proof strategy.

You could also solve the problem by introducing order, and picking 5 bulbs one at a time, without replacement, rather than all 5 at once, and the you would still get the same answer (why? convince yourself). □

Lecture 9: Conditional probability II: breaking complex events into smaller events, methods to solve probability problems, Bayes rule, law of total probability, Bayes theorem

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 18 of the Math for CS book.

1 Breaking complex events into smaller ones

Suppose you have a complex event E whose probability is hard to compute directly. Often times, if you can express E as a union or intersection of simpler events whose probability you know how to compute, then solving the problem becomes much easier. Here are some situations where you could use this approach:

1. Sum rule/inclusion exclusion: Suppose that $E = A_1 \cup \ldots \cup A_n$, where $A_1, \ldots, A_n$ are disjoint, then using the sum rule of probabiltiy we know that
$$P(E) = P(A_1) + \ldots + P(A_n).$$
Now if you know how to compute $P(A_1), \ldots, P(A_n)$ then you are basically done. If $A_1, \ldots, A_n$ are not disjoint, then you have to use inclusion-exclusion which will require you to compute the probabilities of intersections of subsets of $A_1, \ldots, A_n$ which may be possible in certain cases.

2. Chain rule for independent events/general chain rule: The other case is when $E = A_1 \cap \ldots \cap A_n$. We saw how to deal with such situations in the last lecture. When $A_1, \ldots, A_n$ are independent, we know that
$$P(E) = P(A_1)P(A_2)\ldots P(A_n).$$
This is the chain rule for independent events. What if $A_1, \ldots A_n$ are not independent. In the next lecture, we shall see a general chain rule that works in the case when the events are not independent.

The other thing that's often useful is the difference method in probability: often times it's easier to compute $P(E^c)$ and then use $P(E) = 1 - P(E^c)$.

Question . We toss a fair coin 20 times. What is the probability that no two consecutive coin tosses result in the same outcome?

Proof. Let E be the event that no two consecutive coin tosses result in the same outcome. Then, it's not hard to see that:

$$E = \{HTHTHTHTHTHTHTHTHTHT, THTHTHTHTHTHTHTHTHTH\}.$$

Thus, we have that

$$P(E) = P(\{HTHTHTHTHTHTHTHTHTHT\} \cup \{THTHTHTHTHTHTHTHTHTH\})$$

$$= P(\{HTHTHTHTHTHTHTHTHTHT\}) + P(\{THTHTHTHTHTHTHTHTHTH\}),$$

where the last step follows from the application of the sum rule.

Now, let $A_1, \ldots A_{20}$ be such that A_1 is the event that the first coin toss is heads, A_2 is the event that the second coin toss is tails, and so on. Then

$$P(\{HTHTHTHTHTHTHTHTHTHT\}) = P(A_1 \cap A_2 \cap \ldots \cap A_{20}) = P(A_1)P(A_2)\ldots P(A_n) = \frac{1}{2^{20}}.$$

The above equation follows from two facts: (i) $A_1, \ldots, A_{20}$ are independent (since we have 20 independent coin tosses), and (ii) every coin toss is heads or tails with equal probability.

One can similarly show that

$$P(\{THTHTHTHTHTHTHTHTHTH\}) = \frac{1}{2^{20}},$$

and so $P(E) = \frac{2}{2^{20}}$. □

2 Methods to solve probability problems (we've seen so far)

We have seen two methods so far:

1. Direction enumeration: This is the method where you define outcomes and the sample space for the problem, choose an appropriate probability distribution for the sample space, define events, and then compute probabilities/conditional probabilities.
2. Tree method: We discussed this in the previous lecture

Typically, you want to use the latter for most problems (unless the problem explicitly asks you to use another method). In some cases, there might be too much branching or the depth of the tree might be too large, and in those cases it's better to use direction enumeration. For example, recall the bulbs in a box problem from last lecture. Using tree method to solve it can get a bit cumbersome (if you use it naively), and so you want to use direct enumeration in such cases. We will later see that the bulbs in the box problem can be solved using the general chain rule pretty easily.

Sometimes you don't need to use either of the two methods and just have to apply definitions and formulas to get the answer:

Question . A survey was done to study consumption of coke and coffee in NJ. 55% of the participants said that they drink coffee every day, 45% of them said that they drink coke a day, and 70% of them said that they drink at least one of the two ever day. Suppose you choose a random participant of the survey and find out that they drink coke every day. What is the probability that they also drink coffee every day?

Proof. Let A be the event that the randomly chosen participant drinks coffee everyday and B be the event that they drink coke everyday. We want to find $P(A|B)$. We know that $P(A) = 0.55$, $P(B) = 0.45$, and $P(A \cup B) = 0.70$ ($A \cup B$ represents the event "at least one of coffee and coke every day"). This means that using inclusion-exclusion for two events

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.30.$$

Thus,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.30}{0.45} = \frac{2}{3}.$$

□

3 Interpreting probability

This is not really something I want to spend too much time on. We will discuss this very very briefly. Your textbook has a good discussion on this, and you can ask me for other references if you are interested. But here we go anyway.

What does it mean to say "the probability of X happening is ..."? There are two schools of thought that deal with the question. It's important to note that the "math" involving the probabilities (the definitions, axioms, properties, etc.) stays the same in both cases — what changes is how you interpret probabilities.

3.1 Frequentist interpretation of probability

The frequentist school of interpretation says that we can only talk about probabilities of events in the case of experiments that can be repeated multiple times (i.e., multiple trials of the same experiment). For example, consider the experiment of tossing a coin and observing the outcome. You can repeat this experiment as many times as you want. The frequentist school of interpretation says you *can* talk about probabilities of events for this experiment. For example, you can ask "what is the probability of seeing heads?". The frequentist school defines the probability of an event of an experiment as follows:

- Let E be the event of the experiment whose probability you want to compute.
- Repeat the experiment N times where N is very large.
- During each repetition, make a note of whether event E happened or not. Let M be the number of repetitions during which the event E happened (out of the N total repetitions).
- Then the probability of E happening when you perform this experiment is $\frac{M}{N}$.

This school of interpretation, however, doesn't try to define probabilities of the following form: "what's the probability that alien life exists?", "what's the probability that the earth will be destroyed in 10,000 years?", or "what's the probability I'll finish HW5 by midnight?".

3.2 Bayesian interpretation

The Bayesian school of probability thinks of probability as a "degree of belief". For example, consider "what's the probability I'll finish HW5 by midnight?". By skimming through the problems on the HW, looking at your schedule for the rest of the day, and past experience of solving HWs in this course, you can try and assign a probability to the event of you finishing HW5 by midnight. The Bayesian school deems that as acceptable. Similarly, you can assign probabilities to "what's the probability that alien life exists?"[1].

The Bayesian interpretation of probability gives us another way to look at conditional probabilities. Let's go back to the finishing your HW by midnight example. Let's denote the event of finishing your HW by midnight by A. Suppose that you (somehow) came up with an estimate of the probability of A happening, i.e. $P(A)$. Suddenly you get a call from close friend saying that their car broke down and they need you to come and pick them up. Let's call the event of their car breaking down and them asking you for help B. Given that B has happened, do you expect the probability of A, i.e. the probability of you finishing the HW by midnight, to change? It certainly seems plausible given that you will probably have to drive out to where they are and help them. So you now need to "update" your estimate of the probability of A happening *given* that B has happened, and that is precisely the conditional probability $P(A|B)$.

4 Bayes rule

The Bayes rule is a consequence of the definition of conditional probability. Purely in terms of algebraic manipulations and equations, it helps you compute $P(A|B)$ if you know $P(B|A)$ or vice-versa for events A and B in some sample space. There is also a Bayesian way to interpret Bayes rule which I will talk about below. Let's first state what Bayes' rule is. Recall that

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

and

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \implies P(A \cap B) = P(A)P(B|A).$$

Let's substitute the value for $P(A \cap B)$ from the second equation into the first one. Doing this gives us:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}.$$

This last equation is called the Bayes rule. Let's first solve a problem using Bayes rule and then talk about the Bayesian interpretation.

Question . 4 cards are picked from a shuffled deck of 52 cards. What's the probability that at least two aces were picked? What's the probability that at least two aces were picked given that at least one ace was picked?

[1] This is something people have already tried to do: `https://en.wikipedia.org/wiki/Fermi_paradox`

Proof. Let A be the event that at least one ace was picked, and let B be the event that at least two aces were picked. The first question is asking us about $P(B)$, and the second question $P(B|A)$. Let's first find $P(B)$.

The outcomes are possible subsets of the 52 cards of size 4 and so the size of the sample space is $\binom{52}{4}$. Also we will use the uniform distribution on the sample space since each outcome is equally likely (the deck is shuffled). Let B^c be all ways of choosing 4 cards such that the number of aces is either one or zero. There are $\binom{48}{4}$ ways to choose 4 cards such that there are no aces, and $4 \cdot \binom{48}{3}$ ways of choosing four cards such that there is exactly one ace (why?). So

$$|B^c| = \binom{48}{4} + 4\binom{48}{3}.$$

This means that $|B| = \binom{52}{4} - \left(\binom{48}{4} + 4\binom{48}{3}\right)$, and so

$$P(B) = \frac{\binom{52}{4} - \left(\binom{48}{4} + 4\binom{48}{3}\right)}{\binom{52}{4}}.$$

Now to find $P(B|A)$, let's instead first find $P(A|B)$ and then use Bayes rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

But what is $P(A|B)$? If you are told that B has happened, i.e. at least two aces were picked, then certainly at least one card was picked, and so A also happened, and thus $P(A|B) = 1$. This means

$$P(B|A) = \frac{P(B)}{P(A)}.$$

We already know $P(B)$ and so we just need to find $P(A)$ to finish the problem. A^c is the event that no ace was picked, and thus $|A^c| = \binom{48}{4}$. Thus,

$$P(A) = 1 - P(A^c) = 1 - \frac{\binom{48}{4}}{\binom{52}{4}} = \frac{\binom{52}{4} - \binom{48}{4}}{\binom{52}{4}}.$$

Plugging $P(A)$ and $P(B)$ back in to the equation for $P(B|A)$ we get

$$P(B|A) = \frac{\binom{52}{4} - \left(\binom{48}{4} + 4\binom{48}{3}\right)}{\binom{52}{4} - \binom{48}{4}}.$$

□

Here is a Bayesian way of thinking of this. If you plug in the value of the binomial coefficients in the above expressions, we find that $P(B) \approx 0.08$, i.e. there is an 8% chance that at least two of the cards will be aces from among the four. Now you are given the information that A has happened, i.e. at least one of the cards is an ace. Intuitively, with the new information, we should update the probability of B happening, and B is intuitively more likely to happen given the information that A happened. $P(B|A)$ is this "updated" probability (probability of B happening given the information that A happened). In fact, if you work out the numbers above, you will see that $P(B|A) \approx 0.30$! This means that the odds in favor of B did increase in light of this new information!

Bayes rule, then, can be thought of as a way to "update" the probability of B happening when new information (in this case, the information that A happened) is available:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

We will come back to Bayes rule later in this lecture when we study Bayes theorem.

5 Law of total probability

Suppose you have a sample space Ω, and an event B in it. Let's say you want to find $P(B)$ but it's pretty hard to do so directly, so you divide your sample space into disjoint cases (the cases are themselves events), say A_1, A_2, A_3, such that all of them are nonempty, and $A_1 \cup A_2 \cup A_3 = \Omega$. Then we know that $B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B)$, and so using the sum rule:

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B).$$

For many problems, this last step is enough and it's easy to find each of the probabilities in the above expression, but sometimes you not directly told what $P(A_1 \cap B)$ is, but are rather told the probability of A_1 happening, i.e. $P(A_1)$, and the probability of B happening *given* that A_1 happens, i.e. $P(B|A_1)$. In such cases, you can write $P(A_1 \cap B) = P(A_1)P(B|A_1)$. Similarly you write expressions for $P(A_2 \cap B)$ and $P(A_3 \cap B)$. This gives:

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3).$$

This is called the law of total probability. The law of total probability is reason why the tree method works, and in fact, the tree method is nothing but a *diagramatic* representation of the law of total probability! Let's see an example to observe this.

Question . There are three boxes, a box containing two gold coins, a box containing two silver coins and a coin, and a box containing a gold and a silver coin. You pick a box at random (each box is equally likely to be picked), and then pick a random coin from the box such that each coin is equally likely to be picked. You put this coin in your pocket. What is the probability that the coin you pocketed is a gold coin?

Proof. Let B be the event that you pocket a gold coin. Let A_1 be the event that the box with two gold coins is picked, A_2 the event that the box with two silver coins and a gold coin was picked, and A_3 the event that the box with a gold and a silver coin was picked. Let's first solve this problem using the law of total probability. We will then solving it using the tree method and then compare the two to see what really makes the tree method work.

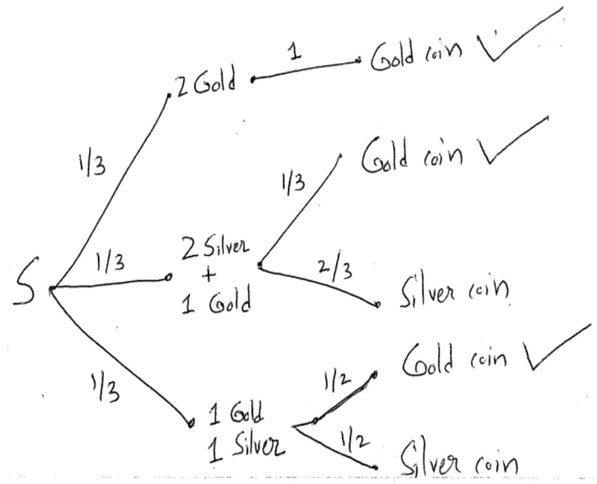
To solve using the law of total probability, let's write down the formula:

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3).$$

Since each box is equally likely to be picked, $P(A_1) = P(A_2) = P(A_3) = 1/3$. What is $P(B|A_1)$? If you pick the box with two gold coins, then when you pick a coin from that box you will *always* pick a gold coin, and so $P(B|A_1) = 1$. Similarly, if you pick the box with two silver and one gold coin, and then pick a coin at random from this box, it's easy to see that the probability of picking a gold coin is 1/3, and so $P(B|A_2) = 1/3$. Using the same idea, we can show that $P(B|A_3) = 1/2$. This means

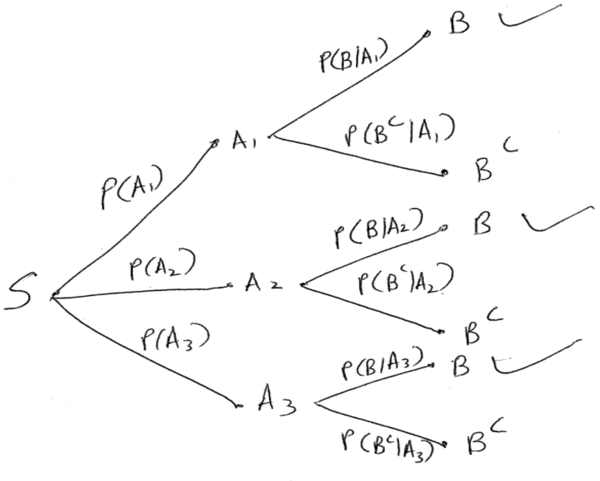
$$P(B) = \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{11}{18}.$$

Let's now solve this problem using the tree method. If you follow the tree method properly, you should end up with (something similar to) the following tree:



Next you would find the probabilities for each of the checked leaves by multiplying the numbers along the root to leaf path, and then add up the probabilities of all the checked leaves. This will give you $\frac{11}{18}$, which is the same answer as we got above.

You might wonder why the tree method works. The best way to see this is to redraw the tree but this time instead of putting in actual numbers and event descriptions, let's use the notation we set up, (i.e., A, B, $P(B|A_1)$, $P(A_1)$, etc.):



If you now multiply the probabilities along the root to leaf path for every checked leaves, and then add those expressions up, you exactly end up with the equation for the law of total probability:

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3).$$

This should give you some insight into why the tree method works: it's just a diagramatic representation of the law of total probabilities, in some sense. □

6 Bayes theorem

Bayes theorem is what you get when you combine Bayes rule with the law of total probability. To motivate Bayes theorem, let go back to the example we dealt with in the previous section.

Question . Recall that in the first step of the experiment, we randomly choose one of the three boxes such that each box is equally likely to be picked. What is the probability that the box with 2 silver coins and 1 gold coin will be picked? Well, that's easy: it's 1/3, i.e. $P(A_2) = 1/3$. Now imagine the following scenario: while you were away you friend performed the experiment from start to finish, and later when you meet her for dinner she tells you that she pocketed a gold coin at the end of the experiment but doesn't remember anything else about how the experiment proceeded! How likely is it that she chose box 2 (the one with 2 silver and 1 one gold coin) in step 1? Is it still 1/3?

Proof. Basically, the problem is asking us to compute $P(A_2|B)$, i.e. the probability that the second box was picked *given* that a gold coin was pocketed in the last step, and asking us to compare $P(A_2|B)$ with $P(A_2)$. To compute $P(A_2|B)$ we can use Bayes rule:

$$P(A_2|B) = \frac{P(B|A_2)P(A_2)}{P(B)}.$$

We saw in the previous section that using the law of total probability/tree method:

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3).$$

Substituting the expression for $P(B)$ back into the Bayes rule equation, we get

$$P(A_2|B) = \frac{P(B|A_2)P(A_2)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)}.$$

This last equation is called the *Bayes theorem.*

Let us finish this problem. If you look at the tree, then the root-to-leaf path that corresponds to the expression $P(B|A_2)P(A_2)$ is the one that starts at S, then goes through "2 Silver + 1 Gold", and finally ends up at "Gold coin". The probability of following this path is $\frac{1}{3} \cdot \frac{1}{3} = 1/9$, and thus:

$$P(A_2|B) = \frac{P(B|A_2)P(A_2)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)} = \frac{\frac{1}{9}}{\frac{11}{18}} = \frac{2}{11}.$$

This means given the information that a gold coin was pocketed, the probability of that the second bad (with 2 silver and 1 gold coin) was picked has gone down (relative to the probability of the same event when no such information was provided), i.e. $P(A_2|B) < P(A_2)$. This makes intuitive sense (can you argue why?). □

The general form of Bayes theorem is the following:

Theorem 1. *Let $A_1, \ldots, A_n$ be nonempty disjoint events of a sample space Ω such that $\bigcup_{i=1}^n A_i = \Omega$*[2]*. For any nonempty event B, and any $1 \leq i \leq n$:*

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^n P(A_j)P(B|A_j)}.$$

Proving this general statement is a problem in Homework 5.

One way to think of Bayes theorem is as follows. The event B can take place in many ways: either A_1 first happens with probability $P(A_1)$ and then B happens with probability $P(B|A_1)$, or A_2 happens first with probability $P(A_2)$ and then B happens with probability $P(B|A_2)$, and so on. So there are n events/cases "through which" B can happen: $A_1, \ldots, A_n$, and furthermore there are none others. The law of total probability, then, lets you compute the probability of B happening. The Bayes theorem on the other hand let's you compute the probability that A_i happened given that B happened, or the probability that B happened "via" A_i given that B did happen.

It's like thinking of $A_1, \ldots, A_n$ as different paths to reach the destination B. You choose path A_i with probability $P(A_i)$, and once you have chosen A_i, you are guaranteed to reach the destination B with probability $P(B|A_i)$, and there is some probability, i.e. $P(B^c|A_i)$ that your vehicle may break down while you are on path A_i and you may never reach B. So the total probability of reaching B is

$$P(B) = \sum_{j=1}^n P(A_j)P(B|A_j),$$

and, given the information that you did reach B, the probability that you reached B by taking path A_i is exactly the expression of Bayes theorem:

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^n P(A_j)P(B|A_j)}.$$

Let us see another example that uses Bayes theorem.

Question . 1% of the women in the age group of $40-50$ have breast cancer. When used on a woman who has breast cancer, a mammogram test will be positive 90% of the time, and when used on someone who doesn't have breast cancer, the test will erroneously result in a positive 10% of the time. A woman in chosen uniformly at random from the population of women in the age group $40-50$ and is made to go through a mammogram test. If the result of the test is positive, what is the probability that she has breast cancer?

Proof. The experiment is here choosing a random woman in the age group of $40-50$ and making her go through the mammogram test. Let B be the event that she tests positive when tested using the mammogram test. Let A be the event that the woman who was chosen is from among those who have breast cancer. Then A^c is the event that she was chosen from the group that doesn't have breast cancer. We know that $P(A) = 0.01$ and $P(A^c) = 0.99$. We also know that if the test is

[2] In fact, the theorem works for a weaker condition: we only need that $(B \cap A_1) \cup \ldots (B \cap A_n) = B$. Can you explain why?

used on someone who does have breast cancer then the probability that it will result in a positive is 0.90, i.e.

$$P(B|A) = 0.90,$$

and that if it used on someone who doesn't have breast cancer, the probability that it will be test positive is 0.10, i.e.

$$P(B|A^c) = 0.10.$$

We want to find the probability that she has breast cancer *given* that she tested positive, i.e. $P(A|B)$. Cleary $A \cup A^c$ covers the entire population of women in the age group $40 - 50$. So using Bayes theorem we have that

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)} = \frac{0.01 \times 0.90}{0.01 \times 0.90 + 0.99 \times 0.10} = 0.08333333\ldots.$$

Thus, given the information that she tested positive, the probability that she has breast cancer is only 8.33%![3]. □

[3] I recommend reading 18.4.2 from the textbook in your spare time to understand what's going on – I mean if a person tested positive and the test is 90% accurate, shouldn't the probabiltiy of them having the disease be much much higher?

Lecture 10 (Part 1): Conditional probability III: conditional independence, the general chain rule, the birthday paradox

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 19 of the Math for CS book.

One of the things that we will be using in this lecture is the idea of conditioning on more than one events. Say we have three events A, B and C in a sample space Ω equipped with a probability distribution P. Then by $P(A|B, C)$ we basically mean $P(A|B \cap C)$, i.e. the probability of A happening given that *both* B and C have happened. It's not hard to verify the following identity (convince yourself!):

$$P(A|B, C) = \frac{P(A \cap B|C)}{P(B|C)}.$$

(You can verify it by plugging in the expressions for each of the terms on the LHS and RHS).

Obviously, we can condition on more than 2 events, for example given $A_1, \ldots, A_n$, we can talk about $P(A_i|A_1, \ldots, A_{i-1})$, i.e. the probability of A_i happening *given* that all of $A_1, \ldots, A_{i-1}$ have taken place.

1 Conditional independence

Definition 1 (Conditional independence)**.** *Given three events A, B and C in a sample space Ω with probability function P defined on it, A and B are said to be conditionally independent with respect to C if*

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

What conditional independence means is that once we are given the information that C has happened (and our sample space now "shrinks down" to C), A and B behave like independent events.

We will now see that independence and conditional independence are two different things which typically have nothing to do with each other.

Question . Consider the experiment of tossing a coin twice. Let A be the event of observing heads in the first toss and B be the event of observing tails in the second toss. If C is the event that an even number of heads are observed, prove that A and B are independent but not conditionally independent with respect to C.

Proof. It is clear that A and B are independent events (why?). Let's show that $P(A \cap B|C) \neq P(A|C)P(B|C)$ to show that A and B are not conditionally independent with respect to C. Note that the event C corresponds to the outcomes $\{HH, TT\}$, A corresponds to $\{HT, HH\}$, and B corresponds to $\{HT, TT\}$. From this we can see that $P(A|C) = P(B|C) = \frac{1}{2}$ (convince yourself!), however $P(A \cap B|C)$ is zero. □

This means that independence does not imply conditional independence. Let's see an example where we see two events that are not independent but are conditionally independent with respect to another event.

Question . Consider the following experiment. We have a box with two coins, one which is fair, and another which has two heads (we shall call it a two-headed coin). We first randomly pick one of the two coins from the box, and then toss the coin we twice. Let A be the event of seeing a heads in the first toss, B be the event of seeing a heads in the second toss, and let C be the event that the fair coin is picked in step one. Show that A and B are not independent but are conditionally independent with respect to C.

Proof. Using the tree method or Bayes theorem (left as an exercise), you can show that $P(A) = 3/4$, $P(B) = 3/4$, but $P(A \cap B) = 5/8$, and hence A and B are not independent.

Now let us condition on C, i.e. assume that C has happened and a fair coin was picked in step one. Then $P(A|C) = 1/2$, $P(B|C) = 1/2$, and $P(A \cap B|C) = 1/4$ (why?), and so A and B are conditionally independent with respect to C. □

2 General chain rule

We saw the chain rule for independent events: if $A_1, \ldots, A_n$ are mutually independent events then $P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n)$. But what can we say if $A_1, \ldots, A_n$ are not mutually independent?

One way to understand this case is the following. We are interested in the probability that all of $A_1, \ldots, A_n$ happen. Let's try to compute it step-by-step or in stages, such that

- First A_1 happens with probability $P(A_1)$.
- Then *given* that A_1 has taken place, A_2 happens with probability $P(A_2|A_1)$. So the probability of A_1 and A_2 happening in this order is $P(A_1)P(A_2|A_1)$ (This can also be proved using the definition of conditional probability).
- Next, given that A_1 and A_2 have happened, A_3 happens with probability $P(A_3|A_1, A_2)$. So the probability of A_1, A_2, and A_3 happening in that order is $P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)$.

$$\vdots$$

$$\vdots$$

- And finally, A_n happens given that $A_1, \ldots, A_{n-1}$ have happened, and the probability of this is $P(A_n|A_1, \ldots, A_{n-1})$, and so the probability of $A_1, \ldots, A_n$ happening, one-by-one, in that order is

  $$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)P(A_4|A_1, A_2, A_3)\ldots P(A_n|A_1, \ldots, A_{n-1}).$$

  This is called the (general) chain rule.

It's not that hard to formally prove the general chain rule using induction and the definition of conditional probability. We will skip the proof here and move straight to an important example that employs the chain rule.

3 Birthday paradox

Let us make some simplifying assumptions before state the problems:

1. There are always 365 days in an year, i.e. assume that there are no leap years. Furthermore, let's number the days as 1 to 365, where day 1 is January $1st$ and day 365 is December 31st.
2. Everyone is/was born in a non-leap year.
3. If we pick a random person, then they are equally likely to have been born on any of the 365 days (let's forget about the year of birth).

Question . Assume the above conditions. Now suppose we choose n random persons to form a group. What is the smallest value of n for which we will *definitely* have at least two people in the group who share a birthday (disregarding the year of course)? What is the smallest value of n so that with *probability* 99% we have at least two people in the group who share a birthday?

Proof. If we wanted to be a 100% sure that two people in the grop have the same birthday, we would have to have a group of at least 366 people (why?), and so $n = 366$ is the smallest value we can go for.

Here is a paradox: if you are willing to go from 100% sure to 99% sure, a much much smaller value of n can do the job, i.e. for a significantly smaller value of n (as compared to 366), we can be 99% sure that two people in the randomly formed group have the same birthday.

Let A denote the event that at least two people in the group have the same birthday. We will find $P(A^c)$. Let us assume that we are choosing the n people for the group, one at a time from the population, at random. Let A_1 be the event that the first person has a birthday on one of the 365 days (obviously, $P(A) = 1$ based on our assumptions). Let A_2 be the event that the first two persons that are chosen have distinct birthdays, let A_3 be the event that the first three people that are chosen have distinct birthdays, $\ldots$, A_i be the event that the first i people that are chosen have all distinct birthadys, $\ldots$, and let A_n be the event that the all n people that were chosen have distinct birthdays. Then

$$A^c = A_1 \cap A_2 \cap \ldots \cap A_n.$$

So basically, we have to find $P(A_1 \cap \ldots \cap A_n)$. Are $A_1, \ldots, A_n$ independent events? They are not (why?), and so we will use the general chain rule to find this probability. Recall that the general chain rule says that

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)\ldots P(A_n|A_1, A_2, \ldots, A_{n-1}).$$

Let's try to compute each of the probabilities involved in the above expression. What's $P(A_1)$? Well that's just 1 and we discussed this above. What about $P(A_2|A_1)$? We want that when we choose the second person, their birthay should be distinct from that of the first person, and the probability of that is $1-\frac{1}{365}$ (why?). Next, given that the first two people have distinct birthdays, what's the probablity that the third person has a birthday distinct from the first two? That's $1-\frac{2}{365}$, and so $P(A_3|A_1, A_2) = 1-\frac{2}{365}$. Following the same idea, we can conclude that the probability that the i$^{\text{th}}$ person has a birthday distinct from the first $i-1$ people given that the first $i-1$ people have all distinct birthdays is $1-\frac{(i-1)}{365}$, i.e. $P(A_i|A_1,\ldots,A_{i-1}) = \frac{365-(i-1)}{365}$. Thus,

$$P(A_1\cap\ldots\cap A_n) = 1(1-\frac{1}{365})(1-\frac{2}{365})\ldots(1-\frac{n-1}{365}).$$

We will now use a basic fact from calculus:

Fact 2. *For all $x \geq 0$,*

$$(1-x) \leq e^{-x}.$$

Let's apply this inequality to each of the terms, i.e. use the fact that for every $2 \leq i \leq n$,

$$1-\frac{(i-1)}{365} \leq e^{-\frac{(i-1)}{365}}.$$

So we get that

$$P(A_1\cap\ldots\cap A_n) \leq e^{-\frac{1}{365}}e^{-\frac{2}{365}}\ldots e^{-\frac{(n-1)}{365}} = e^{-\frac{(1+2+3+\ldots+(n-1))}{365}} = e^{-\frac{n(n-1)}{730}},$$

where the last equality uses the fact that $1+2+\ldots+(n-1) = \frac{n(n-1)}{2}$.

All this means that the probability that all n people have distinct birthdays is

$$P(A^c) = P(A_1\cap\ldots\cap A_n) \leq e^{-\frac{n(n-1)}{730}}.$$

We want the probability that at least two people have the same birthday to be ≤ 0.99, i.e. $P(A) \geq 0.99$, and so we want that $P(A^c) \leq 0.01$. This basically means we want to choose a n that's large enough so that $P(A^c) \leq 0.01$. To find such an n, recall that

$$P(A^c) \leq e^{-\frac{n(n-1)}{730}}.$$

So if we can choose n so that the RHS becomes less than 0.01, we should be good. So we want

$$e^{-\frac{n(n-1)}{730}} \leq 0.01.$$

Inverting both sides, and flipping the direction of the inequality, we get

$$e^{\frac{n(n-1)}{730}} \geq 100$$

We can now take natural logarithm on both sides, to get

$$\frac{n(n-1)}{730} \geq \ln(100)$$

or

$$n(n-1) \geq 730\times\ln(100).$$

If we can ensure that $(n-1)(n-1)$ is larger than the RHS than surely $n(n-1)$ is also larger than the right hand side, since $n(n-1) \geq (n-1)(n-1)$. So let's try to find an n such that

$$(n-1)(n-1) = (n-1)^2 \geq \ln(100)\times 730.$$

or

$$n \geq \sqrt{730 \times \ln(100)} + 1.$$

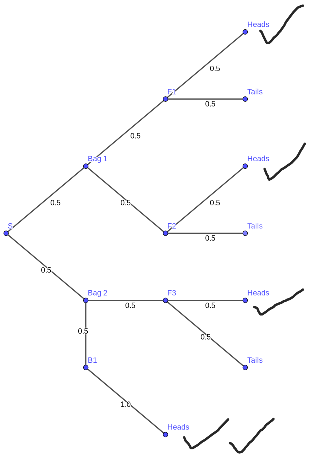
The RHS works out to be about 58. This means that if we choose an n that's at least 58, we can ensure that $P(A^c) \leq 0.01$. Let's choose $n = 60$. This is incredible!! If we willing to settle for 99% sureity instead of being a 100% sure, we need only 60 randomly chosen people to guarantee that two people have the same birthday! This is a sixth of 365, the number of people needed to ensure that you always have two people with the same birthday. □

4 The general chain rule and the tree method

We saw informally in the previous lecture that the "engine" behind the tree method is basically the law of total probability. Well, I lied a bit. While the version of the law of total probability we saw in the last lecture does explain why the tree method works for cases when the *depth* of the tree is two, it doesn't fully explain the case when the depth of the tree is more than two. By depth here I mean the length of the longest root to leaf path. One needs the general chain rule to address the case when the depth is more than two. Let's consider an example from Lecture 8 to demonstrate this.

Question . I have two bags. Bag 1 contains two fair coins and Bag 2 contains a fair coin and a coin with two heads. I toss a fair coin. If the outcome is heads, I pick Bag 1, and if the outcome is tails I pick Bag 2. Then from the bag in picked I the previous step, I randomly pick a coin such that both coins in the bag are equally likely to be picked. Finally I toss the coin I picked from the bag. What is the probability that the final outcome is heads?

Proof. Here is the tree (F_1 and F_2 are the two fairs coin in Bag 1, and F_3 and B_1 are the fair and two-headed coins respectively in Bag 2):



You will notice that the tree has depth/height 3: the first node branches into two possibilities depending on which box is picked, then the next level of nodes branch into two possibilities each depending which coin is picked, and finally there is another branching based on what the coin toss results in.

Consider the leaf that corresponds to the event of picking the box with a fair coin and a two-headed coin, then picking the fair coin, and then seeing a heads (this is the third leaf from the top with a check mark). The way we calculate the probability of this event is by multiplying all the probabilities along the root to leaf path for this event. Why does the multiplication work? Let's try to understand that now:

Let A_1 be the event of picking the box with a fair coin and a two-headed coin, let A_2 be the event of picking the fair coin, and let A_3 be the event of seeing a heads in the final toss. The leaf we talked about in the previous paragraph basically corresponds to $A_1 \cap A_2 \cap A_3$, and as discussed above the way we find the $P(A_1 \cap A_2 \cap A_3)$ is by multiplying the root-to-leaf probabilities. To see why the multiplication works, note that the $P(A_1)$ is 0.5. Given that A_1 has happened, i.e. the box with a fair coin and a two-headed coin is picked, the probability of picking a fair coin is 0.5, i.e. $P(A_2|A_1) = 0.5$. Finally, given that Box 2 was picked *and* the fair coin was picked from it, it's clear that the probability of seeing a heads is 0.5, i.e. $P(A_3|A_1, A_2) = 0.5$. Using the general chain rule, we know that

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) = 0.5 \times 0.5 \times 0.5.$$

So the reason why we multiply is because we are implicitly using the general chain rule! Each of the branches is basically labelled with a conditional probability!

Now we proceed in a similar way: we multiply all the probabilities on the root-to-leaf path for every leaf with a check mark, and the reason we do this is the same as we discussed above: we are using the general chain rule! Once we compute all the needed probabilities, we add them up to get the total probabilty of seeing heads. The last step of adding these probabilities is similar to the law of total probability.

Basically, what we are doing is, to *partition* A_3, the probability of seeing a heads, into various disjoint cases: "Box 1 and fair coin and heads", "Box 1 and fair coin and tails", "Box 2 and fair coin and heads",..., etc. More formally,

$$P(A_3) = P(A_1 \cap A_2 \cap A_3) + P(A_1 \cap A_2^c \cap A_3) + P(A_1^c \cap A_2 \cap A_3) + P(A_1^c \cap A_2^c \cap A_3).$$

(Convince yourself that all these sets form a partition of A_3). The four probabilities in the above expression correspond to the four leaves with a check mark. This explains why we add them up. Each of these probabilities can then be computed using the general chain rule as discussed above. □

Lecture 10 (Part 2): Random variables I: definition, events and random variables, independent random variables

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 19 of the Math for CS book.

1 Introduction to random variables

Recall the steps involved in modelling random experiments: define outcomes in a suitable manner and this defines Ω the sample space and define an appropriate probability distribution on Ω. Ω and P together are said to define a *probability space*. We denote the probability space by (Ω, P).

Typically, in every random experiment, we are interested in "measuring" some quantity that depends on the outcome of the experiment. Informally, speaking these quantities of interest that are measured during an experiment are called random variables.

For example, consider the experiment of choosing a random person from the population of adults in the age group $18 - 25$ in the US (say at the present moment) and then measuring their height. Clearly, height is a quantity that depends on the outcome of the experiment (i.e., it depends on which person was chosen randomly during the experiment). In this case, "heigh" is a random variable associated with the experiment.

Definition 1. *Let (Ω, P) be a probability space then a random variable X is a function with domain Ω and codomain $\mathbb{R}$, i.e. $X : \Omega \to \mathbb{R}$.*

Let's look at some more examples. Consider the experiment of tossing a fair coin 3 times and let X be the random variable that is equal to the number of heads observed during the 3 tosses. For this experiment, the sample space is nothing but length 3 sequences made from the letters H and T. Since we are tossing a fair coin, we define the probability distribution on P to be the uniform distribution.

X is a function from Ω to $\mathbb{R}$. Let's try to see what value X takes on various outcomes. Recall that

$$\Omega = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}.$$

Then,

$$X(TTT) = 0$$
$$X(TTH) = X(THT) = X(HTT) = 1,$$
$$X(THH) = X(HTH) = X(HHT) = 2,$$
$$X(HHH) = 3.$$

For a function $f : C \to D$, recall that the range of f is all those values in D which the function f can take, i.e.

$$Range(f) = \{d \in D |\ \exists c \in C,\ f(c) = d\}.$$

For example, consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$, then the codomain of f as it is defined is $\mathbb{R}$ but the range consists of all positive real numbers (since the square of a real number can never be negative), i.e. $Range(f) = \{x \in \mathbb{R}|\ x > 0\}$.

The codomain of the random variable X in the above example is $\mathbb{R}$ but $Range(X) = \{0, 1, 2, 3\}$.

Here is another example. Suppose you roll a dice, and let X be the number rolled by the dice. Then X is a random variable with $Range(X) = \{1, 2, \ldots, 6\}$.

Here is one last example which, as we shall see in the next section, is of particular interest. Consider the experiment of tossing a coin 100 times. Let Y be the random variable that is 1 if all the coin tosses result in the same outcome, and is 0 otherwise. In this case, the range of Y is just $\{0, 1\}$. Let E denote the event that all the tosses result in the same outcome, then it's almost as if Y is a "bulb" (i.e., Y becomes 1) that turns on at the end of the experiment if event E happens, and doesn't turn on (i.e., $Y = 0$) if E doesn't happen.

2 Random variables and events

As we will see, events and random variables have a lot more in common than what one might expect at first. We will show how to use events to define random variables, and then show how random variables naturally define events.

2.1 From events to random variables

Definition 2 (Indicator random variable)**.** *Let E be an event of a probability space (Ω, P). Define the random variable X_E as follows: for every outcome ω of the sample space Ω,*

$$X_E(\omega) = 1 \textit{ if } \omega \textit{ is an outcome in } E$$

$$X_E(\omega) = 0 \textit{ if } \omega \textit{ is NOT an outcome in } E$$

That is X_E takes value 1 if the event E happens, and takes the value 0 if E doesn't happen. X_E is called the indicator random variable for the event E (or associated with the event E).

There are different notations for representing the indicator random variable of an event E: I_E, $I[E]$, etc., but I don't mind you using whatever notation you like as long as you explicitly mention that the variable in question is an indicator random variable!

So if we go back to the example of tossing 100 coins, then the random variable X in that case was actually an indicator random variable for the event of all 100 tosses resulting in the same outcome, i.e. E is the event consisting of the all heads outcome and the all tails outcome.

2.2 From random variables to events

Definition 3. *Let X be a random variable defined on a probability space (Ω, P) and let $c \in \mathbb{R}$ be some real number, then $[X = c]$ denotes the event that X takes the value c, or more formally*

$$[X = c] = \{\omega \in \Omega | \; X(\omega) = c\},$$

that is the set of all outcomes on which X takes value c.

Again, even though I prefer if you use the notation $[X = c]$ to denote such events, it's okay if you use other notation as long as you make it clear that the event in question is the one that corresponds to X taking value c. Another thing to note here is that $[X = c]$ is an event, it's a subset of Ω, and consists of outcomes for which $X = c$.

If we go back to the example of tossing three coins and letting X be the random variable that counts the number of heads observed, then the event $[X = 2]$ is give by

$$[X = 2] = \{HHT, HTH, THH\}.$$

Obviously, we are not restricted to using equalities involving random variables to define events. I can also define event $[X > 1]$, i.e. the event of X taking a value larger than 1. In the three coin tosses example,

$$[X > 1] = \{HHT, HTH, THH, HHH\}.$$

2.3 Probability and random variables

Having seen this duality between random variables and events, we can now ask questions of the following form: "In the three coin toss example, what is the probability that $0 < X \leq 2$?".

Such questions can easily be answered: if X is a random variable defined on a probability space (Ω, P) then the probability of X taking a value between a and b is basically the probability of the event $[a \leq X \leq b]$, i.e. $P([a \leq X \leq b])$. We will drop the square brackets in such cases and simply write

$$P(a \leq X \leq b).$$

For example, in the three coin tosses example, $P(0 < X \leq 2)$ is just 3/4 (why?). Here are some more examples.

Question . A fair coin is tosses 100 times. Let X be the number of heads observed. What is $P(X > 1)$?

Proof. We are interested in the probability of the event $[X > 1]$. Note that the range of X is $\{0, 1, \ldots, 100\}$. So the complement of the event $[X > 1]$ is $[0 \leq X \leq 1]$. In fact, $[0 \leq X \leq 1]$ can be written as the union of two disjoint events:

$$[0 \leq X \leq 1] = [X = 0] \cup [X = 1].$$

(Convince yourself!) $[X = 0]$ is the event that none of the 100 coin tosses result in heads and thus

$$P(X = 0) = \frac{1}{2^{100}},$$

while $[X = 1]$ is the event of seeing exactly one heads during the 100 tossees, and so

$$P(X = 1) = \frac{100}{2^{100}}$$

(Can you say why?). Thus, using the sum rule for probability we know that

$$P(0 \leq X \leq 1) = P(X = 0) + P(X = 1) = \frac{101}{2^{100}}.$$

Since $[X > 1]$ is the complement of $[0 \leq X \leq 1]$, we get

$$P(X > 1) = 1 - P(0 \leq X \leq 1) = 1 - \frac{101}{2^{100}}.$$

□

Question . A fair coin is tossed 100 times. Let E be the event of seeing an even number of heads, and let X be the indicator random variable of E. What is $P(X = 0)$?

Proof. Since X is the indicator random variable of E, X is 1 when E happens and is 0 if E doesn't happen. Thus,

$$P(X = 0) = P(\text{E doesn't happen}) = P(\text{Number of heads observed is odd}).$$

Recall that for coin tosses one possible choice of the sample space is the set of all binary strings of a given length. In this case we are tossing a coin 100 times, so we define the sample space as the set of all binary strings of length 100. Also recall from early lectures that the number of binary strings with an even number of ones is the same as the number of binary strings with an odd numbers ones. This means that

$$P(X = 0) = P(\text{Number of heads observed is odd}) = \frac{1}{2}.$$

(It might be a good idea to referesh this if you seemed to have forgotten it) □

3 Independent random variables

Suppose we are sampling a random person from the population of adults in the US in the age group $25 - 30$. Let X be the current annual income of the randomly chosen person, and let Y be the joint income of their parents when they were growing up. Do we expect X and Y to be "correlated", or do we expect these variables to be able to take values "independent" of each other? We expect them to be correlated! For example, if $Y \geq \$1000000$, it is highly unlikely that $X \leq \$40,000$.

On the other hand, if Y was the month they were born in (we can represent the months using numbers from 1 to 12) then we don't expect X and Y to be correlated. We expect the probability of X taking a certain value to remain unchanged even if Y was fixed to some value. We can use this to formalize what it means for two random variables to be independent.

Definition 4. *Let (Ω, P) be a probability space and let X and Y be random variables defined on it. We say that X and Y are independent if, for every $a \in \mathbb{R}, b \in \mathbb{R}$ the events $[X = a]$ and $[Y = b]$ are independent, or in other words, for every $a \in \mathbb{R}$, $b \in \mathbb{R}$,*

$$P(X = a \cap Y = b) = P(X = a)P(Y = b).$$

Note that if some number c is not the range of X then $P(X = c) = 0$, and similarly if some number d is not in the range of Y then $P(Y = d) = 0$, and so it doesn't make sense to consider values that are not in the range of X or Y in the above definition (because for such values both the LHS and RHS of the equation in the above definition are trivially zero). We can instead say that X and Y are independent if and only if for every $a \in \mathit{Range}(X)$ and every $b \in \mathit{Range}(Y)$, $P(X = a \cap Y = b) = P(X = a)P(Y = b)$ (Convince yourself that this is indeed true).

Let's look at some examples. Suppose we consider the tossing 100 coins experiment again, and let X be the number of heads we see, and let Y be the indicator random variable for the event E that all tosses result in the same outcome (all tosses result in heads or all tosses result in tails), then clearly X and Y are not independent: $P(X = 35 \cap Y = 1) = 0$ because if $Y = 1$ then X must be 0 or 100, whereas $P(X = 35) > 0$ and $P(Y = 1) > 0$ and so $P(Y = 1)P(X = 35) > 0$.

Consider the experiment of rolling two dice, one blue and one red, and let X be the random variable that is the number rolled by the red dice and let Y be the number rolled by the blue dice, then X and Y are independent. To see this, note that the range of both X and Y is $\{1, \dots, 6\}$, and for any a such that $a < 1$ or $a > 6$ $P(X = a) = P(Y = a) = 0$, so only need to prove that for every $a, b \in \{1, \dots, 6\}$,

$$P(X = a \cap Y = b) = P(X = a)P(Y = b).$$

To see this is true, let a and b be arbitrary numbers in $\{1, \dots, 6\}$. We know that $P(X = a) = P(Y = b) = 1/6$ (make sure you know why this is true). On the other hand, $P(X = a \cap Y = b) = \frac{1}{36}$ since there is only one outcome where the red dice rolls the number a and the blue dice rolls the number b. Thus, $P(X = a \cap Y = b) = P(X = a)P(Y = b)$.

Question . Consider the experiment of tossing a coin two times, and let X be 1 if the 1^{st} coin toss is heads and 0 otherwise, and let Y be the indicator random variable for the event that we observe an even number of heads. Show that X and Y are independent.

Proof. First notice that the range of X and Y is just $\{0, 1\}$, and so we need to show that for every $a, b \in \{0, 1\}$, $P(X = a \cap Y = b) = P(X = a)P(Y = b)$. Here I will only show you the proof for $a = 0$ and $b = 1$. You can easily complete the proof for the other cases.

We know that $P(X = 0) = 1/2$ and $P(Y = 1) = 1/2$. What about $P(X = 0 \cap Y = 1)$? $[X = 0] \cap [Y = 1]$ is the event consisting of all outcomes where the first coin is tails and there are an even number of heads. The only possible outcome of this form is TT (why?), and thus $P(X = 0 \cap Y = 1) = 1/4$. This show that $P(X = 0 \cap Y = 1) = P(X = 0)P(Y = 1)$. □

Lecture 11: Random variables II: mutual independence, k-wise independence, some obvious facts about random variables, expectation of a random variable, linearity of expectation

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 19 of the Math for CS book.

1 Mutually independent random variables

Definition 1. *Let $X_1, \ldots, X_n$ be random variables defined on a probability space (Ω, P). Then $X_1, \ldots, X_n$ are said to mutually independent if for all $a_1, a_2, \ldots, a_n \in \mathbb{R}$, the events $[X_1 = a_1], [X_2 = a_2], \ldots, [X_n = a_n]$ are mutually independent, or, in other words, for all $a_1, \ldots, a_n \in \mathbb{R}$,*

$$P(X_1 = a_1 \cap X_2 = a_2) = P(X_1 = a_1)P(X_2 = a_2)$$

$$P(X_1 = a_1 \cap X_3 = a_3) = P(X_1 = a_1)P(X_3 = a_3)$$

$$\vdots$$

$$P(X_1 = a_1 \cap X_n = a_n) = P(X_1 = a_1)P(X_n = a_n)$$

$$P(X_2 = a_2 \cap X_3 = a_3) = P(X_2 = a_2)P(X_3 = a_3)$$

$$\vdots$$

$$P(X_{n-1} = a_{n-1} \cap X_n = a_n) = P(X_{n-1} = a_{n-1})P(X_n = a_n)$$

$$P(X_1 = a_1 \cap X_2 = a_2 \cap X_3 = a_3) = P(X_1 = a_1)P(X_2 = a_2)P(X_3 = a_3)$$

$$\vdots$$

$$P(X_{i_1} = a_{i_1} \cap \ldots X_{i_k} = a_{i_k}) = P(X_{i_1} = a_{i_1}) \cdot \ldots \cdot P(X_{i_k} = a_{i_k})$$

$$\vdots$$

$$P(X_1 = a_1 \cap X_2 = a_2 \cap \ldots \cap X_n = a_n) = P(X_1 = a_1) \cdot \ldots \cdot P(X_n = a_n).$$

As in the case of the definition of independence for two random variables, we can restrict our attention to the ranges of the random variables instead of all real numbers in the definition(convince yourself that the two definitions are equivalent): $X_1, \ldots X_n$ are independent if and only if for all $a_1 \in Range(X_1), a_2 \in Range(X_2), \ldots, a_n \in Range(X_n)$, the event $[X_1 = a_1], [X_2 = a_2], \ldots, [X_n = a_n]$ are mutually independent.

For example, if we roll a dice 100 times, and X_i is the number rolled in the i^{th} roll, then variables $X_1, \ldots, X_{100}$ are mutually independent.

An obvious yet useful consequence is the following:

Fact 2. *If $X_1, \ldots, X_n$ are mutually independent random variables defined on a probability space (Ω, P) then for every $a_1 \in Range(X_1), a_2 \in Range(X_2), \ldots, a_n \in Range(X_n)$,*

$$P(X_1 = a_1 \cap X_2 = a_2 \cap \ldots \cap X_n = a_n) = P(X_1 = a_1) \cdot \ldots \cdot P(X_n = a_n).$$

1.1 Detour: a warning about the definition of independence

When are events A, B and C of a probability space (Ω, P) mutually independent? If we know that $P(A \cap B \cap C) = P(A)P(B)P(C)$ can we conclude that they are mutually independent?

The anwer is no! It is possible to have three events A, B and C such that $P(A \cap B \cap C) = P(A)P(B)P(C)$ yet they are not mutually independent (See the example given here: `http://www.engr.mun.ca/~ggeorge/MathGaz04.pdf`). Recall that for the three events to be mutually independent, we need

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

and

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

It is important to keep this in mind!

2 k-wise independence

Ever wonder why we keep referring to more than 2 independent events/random variables as being "mutually independent" and not simply "independent"? This is because there are various kinds of independence possible when we are dealing with three or more events/random variables[1].

Definition 3 (k-wise independent events)**.** *Let $A_1, \ldots, A_n$ (assume $n \geq 3$) be events of a probability space (Ω, P), then $A_1, \ldots, A_n$ are said to be k-wise independent (for $2 \leq k \leq n$) if any k of them are mutually independent, i.e. if we pick k events from $A_1, \ldots, A_n$, say $A_{i_1}, \ldots, A_{i_k}$, then $A_{i_1}, \ldots, A_{i_k}$ are mutually independent.*

Of course, if $A_1, \ldots, A_n$ are n-wise independent then they are basically mutually independent. When $k = 2$, we say $A_1, \ldots, A_n$ are *pairwise* independent events.

All this can be extended naturally to random variables:

Definition 4 (k-wise independent random variables)**.** *Let $X_1, \ldots, X_n$ (assume $n \geq 3$) be random variables defined on a probability space (Ω, P), then $X_1, \ldots, X_n$ are said to be k-wise independent (for $2 \leq k \leq n$) if any k of them are mutually independent, i.e. if we pick k random variables from among $X_1, \ldots, X_n$, say $X_{i_1}, \ldots, X_{i_k}$, then $X_{i_1}, \ldots, X_{i_k}$ are mutually independent.*

When $k = 2$, we say $X_1, \ldots, X_n$ are *pairwise* independent random variables. Let's see an example of three random variables that are pairwise independent but not mutually independent.

Suppose we toss a coin three times. Assume that the coin is fair. Let X be 1 if the first coin toss is heads and 0 otherwise, let Y be 1 if the second coin toss is tails and 0 otherwise, and let Z be

[1] For two random variables, there is only one notion of independence.

the indicator random variable for the event of seeing an even number of heads. Then X and Y are independent (why?), X and Z are independent(why?), and Z and Y are independent (why?). This means that any two out of three random variables are independent. However, X, Y, and Z are not mutually independent. To see this, observe that

$$P(X = 1 \cap Y = 1 \cap Z = 1) = 0 \neq P(X = 1)P(Y = 1)P(Z = 1),$$

which contradicts one of the requirements for them to be independent.

In general, there can be n random variables $X_1, \ldots, X_n$ such that they are k-wise independent for some $1 \leq k \leq n - 1$ but not mutually independent, i.e. to say that k-wise independence (for $1 \leq k \leq n - 1$) is strictly weaker than mutual independence (the latter implies the former).

3 Some obvious stuff

Here are some obvious facts/definitions that follow from everything we have seen so far about probability, events, and random variables. I will state them without proof and it's up to you to verify them (I recommend doing it!). Let's assume that X is a random variable defined on a probability space (Ω, P).

1. For every $a \in \mathbb{R}$, $0 \leq P(X = a) \leq 1)$.
2. If $a \notin Range(X)$, then $P(X = a) = 0$.
3. $P(X \in \mathbb{R}) = 1$.
4. The sample space Ω is partitioned into disjoint events by the random variable. e.g., if $Range(X) = \{a_1, \ldots, a_k\} \subset \mathbb{R}$ then the events $[X = a_1], \ldots [X = a_k]$ are (i) disjoint and (ii)their union covers the whole of Ω, i.e. $\bigcup_{i=1}^{k}[X = a_i] = \Omega$. In general, $\bigcup_{a \in Range(X)}[X = a] = \Omega$.
5. $\sum_{a \in Range(X)} P(X = a) = 1$
6. If $a, b \in \mathbb{R}$ such that $a \neq b$ then $P(X = a \cup X = b) = P(X = a) + P(X = b)$.
7. If I_1 and I_2 are disjoint subsets of $\mathbb{R}$, then $P(X \in I_1 \cup X \in I_2) = P(X \in I_1) + P(X \in I_2)$. If $I_1 \cap I_2 \neq \emptyset$ then
$$P(X \in I_1 \cup X \in I_2) = P(X \in I_1) + P(X \in I_2) - P(X \in I_1 \cap I_2).$$
8. $P(X \geq a) = 1 - P(X < a)$
9. Let Y be another random variable define on the same probability space. Then if $b \in Range(Y)$ and $a \in Range(X)$, the probability of X taking the value a given that Y takes the value b is
$$P(X = a|Y = b) = \frac{P(X = a \cap Y = b)}{P(Y = b)}.$$
10. X and Y are independent if and only if for every $a \in Range(X)$ and every $b \in Range(Y)$, $P(X = a|Y = b) = P(X = a)$.

4 Expectation of random variables

Consider the set of all students in the summer 2018 CS 206 class. Let this be our sample space Ω. Furthermore, let us assume that we picking a student at random so that every student is equally likely to be picked, and thus we use the uniform distribution on Ω. Let X the following random variable defined on Ω: for every student $\omega \in \Omega$, $X(\omega)$ is the midterm score of the student ω. Then, the average midterm score of the class is

$$\frac{\sum_{\omega\in\Omega} X(\omega)}{|\Omega|} = \sum_{\omega\in\Omega} X(\omega)\frac{1}{|\Omega|} = \sum_{\omega\in\Omega} X(\omega)P(\{\omega\}).$$

(Can you see why all these equalities are true?). The expression $\sum_{\omega\in\Omega} X(\omega)P(\{\omega\})$ is called the *expected value* of the random variable X, and is denoted by $\mathbb{E}[X]$, i.e.

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega)P(\{\omega\}).$$

The definition is the same even if we were dealing with a random variable X defined on some arbitrary probability space (Ω, P) where P is not necessarily the uniform distribution on Ω, and is some arbitrary probability distribution on Ω. Even in this general case, one can think of the $\mathbb{E}[X]$ as the "average value" of the random variable X.

The expected value of a random variable provides crucial information about the behavior of the random variable, i.e. what values does the random variable typically take and with what probability. Let's see some examples.

Question . Suppos we toss two fair coins and let X be the number of heads observed. What is $\mathbb{E}[X]$?

Proof. There are four outcomes, i.e. $\Omega = \{HH, HT, TH, TT\}$ and we have the uniform distribution on Ω. Furthermore, the random variable X is defined as follows:

$$X(HH) = 2,\ X(HT) = 1,\ X(TH) = 1,\ X(TT) = 0.$$

Then, using the above formula for $\mathbb{E}[X]$ we have

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega)P(\{\omega\}) = X(HH)P(\{HH\})+X(HT)P(\{HT\})+X(TH)P(\{TH\})+X(TT)P(\{TT\})$$

$$= 2\cdot\frac{1}{4} + 1\cdot\frac{1}{4} + 1\cdot\frac{1}{4} + 0\cdot\frac{1}{4} = 1.$$

□

4.1 An alternate definition for expectation

Let us go back to the setup with Ω being the set of all students, a student being randomly chosen, and X being the midterm score of the randomly picked student. Then we saw that the average of the class was basically $\mathbb{E}[X]$. Here is another way to calculate the average of the class. Let's assume that the distinct scores scored by the students on the midterm are $a_1, \dots, a_k$, i.e. the range of X is $\{a_1, \dots, a_k\}$. Let's collect all the students who scored a_1 points. Note that this is basically the event/set $[X = a_1]$. Similarly, we can collect all the students who scored a_2 and the set of those students is basically $[X = a_2]$, and we can repeat the same for $a_3, \dots, a_k$. Thus, we have partitioned the students by their scores into k groups (obviously, the groups are disjoint). Another way to compute the average is as follows (convince yourself):

$$\frac{\sum_{i=1}^{k}(a_i \cdot |[X = a_i]|)}{|\Omega|} = \sum_{i=1}^{k} a_i \frac{|[X = a_i]|}{|\Omega|} = \sum_{i=1}^{k} a_i P(X = a_i) = \sum_{a \in Range(X)} aP(X = a).$$

Thus, in this case,

$$\mathbb{E}[X] = \sum_{a \in Range(X)} aP(X = a).$$

Turns out that one can prove that this is true for all random variables defined on arbitrary probability spaces (Ω, P), and so this is a general formula for the expectation.

Let's see an example.

Question . Suppose we toss two dice, and let X be the sum of the numbers rolled by the two dice. Find $\mathbb{E}[X]$.

Proof. We will use the second definition for expectation for this problem. To do so, first notice that the range of X is $\{2, \dots, 12\}$, and we need to find $P(X = 2), P(X = 3), \dots, P(X = 12)$ and then use those values in the above formula. It's not hard to check that (I encourage you to verify this):

$$P(X = 2) = P(X = 12) = \frac{1}{36},$$

$$P(X = 3) = P(X = 11) = \frac{2}{36},$$

$$P(X = 4) = P(X = 10) = \frac{3}{36},$$

$$P(X = 5) = P(X = 9) = \frac{4}{36},$$

$$P(X = 6) = P(X = 8) = \frac{5}{36},$$

$$P(X = 7) = \frac{6}{36}.$$

Thus, using the alternate definition of expectation, we have

$$\mathbb{E}[X] = \sum_{a \in Range(X)} aP(X = a) = \sum_{i=2}^{12} i \cdot P(X = i)$$

$$= 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + \ldots + 7 \cdot \frac{6}{36} + \ldots + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7.$$

□

If you thought the solution to the last problem was a little cumbersome, how about you try this:

Question . Suppose we toss 100 dice, and let X be the sum of the numbers rolled by the 100 dice. Find $\mathbb{E}[X]$.

Obviously, following the same strategy as before might take too long. It turns out that expectation of random variables has a powerful property that makes such problems almost trivial to solve!

4.2 Linearity of expectation

Theorem 5 (Linearity of expectation)**.** *Let X and Y be random variables defined on a probability space (Ω, P). Then,*

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Notice that we do not require any conditions on X and Y, and most surprisingly, then don't need to be independent!

Before we prove the linearity of expectation, let's go back to the problem of rolling 100 dice. Let X_1 be the number rolled by the first dice, X_2 the number rolled by the second dice, …, X_{100} be the number rolled by the 100^{th} dice. Then X, the sum of the numbers rolled by all the dice, is simply $X_1 + X_2 + \ldots + X_n$. Using linearity of expectation, we get

$$\mathbb{E}[X] = \sum_{i=1}^{100} \mathbb{E}[X_i].$$

What is $\mathbb{E}[X_i]$ for some i? This is easy to calculate (convince yourself):

$$\mathbb{E}[X_i] = \frac{1}{6} + 2 \cdot \frac{1}{6} + \ldots + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5,$$

and thus $\mathbb{E}[X] = 350$. Observe how much simpler this solution is compared to the approach we used earlier.

We will look at another nontrivial application of the linearity of expectation.

Question . Suppose there are 100 rabbits $R_1, \ldots, R_{100}$ and 100 rabbit holes $H_1, \ldots, H_{100}$ such that, to begin with, R_i is in H_i. All the 100 rabbits come out of their holes and are hopping around till they see an eagle swooping down. This makes them go helter-skelter and then run back into the holes so that (i) each rabbit is equally likely to go into any of the 100 holes and (ii) no two rabbits go into the same hole. If X is the number of rabbits that go back into their own hole, what is $\mathbb{E}[X]$?

Proof. Remember that we dealt with this problem when we studided inclusion-exclusion. If you go back and look at the notes from that lecture, you will realize that simply computing the expectation using the formula $\mathbb{E}[X] = \sum_{i=1}^{100} i \cdot P(X = i)$ will be really cumbersome, since even just computing $P(X = 0)$ needs the use of inclusion-exclusion and turns out to be not so easy (imagine doing it for $P(X = 1), P(X = 2), \ldots, P(X = 100)$). It can certainly be done but there is a much much simpler way as we shall see now.

Let X_i be the indicator random variable for the event of R_i going back into its own hole H_i. Then, the total number of rabbits that go back into their own hole is simply $X = X_1 + X_2 + \ldots + X_n$, and thus, using the linearity of expectation, we have that

$$\mathbb{E}[X] = \sum_{i=1}^{100} \mathbb{E}[X_i].$$

Let's compute $\mathbb{E}[X_i]$. We know that the range of X_i is just $\{0, 1\}$, and that

$$P(X = 1) = P(R_i \text{ goes back into } H_i).$$

Also, $P(X = 0) = 1 - P(X = 1)$ (why?). Then, using the definition of expectation of X_i we have

$$\mathbb{E}[X_i] = 1 \cdot P(R_i \text{ goes back into } H_i) + 0 \cdot (1 - P(R_i \text{ goes back into } H_i)) = P(R_i \text{ goes back into } H_i)$$

All we need now is to compute $P(R_i \text{ goes back into } H_i)$. But this is super-easy! Since we are told that each rabbit is equally likely to go into any of the 100 holes[2], we have that

$$P(R_i \text{ goes back into } H_i) = \frac{1}{100}.$$

(Convince yourself!) This means that

$$\mathbb{E}[X] = \sum_{i=1}^{100} \frac{1}{100} = 1.$$

Thus, on an average, only one rabbit goes into its own hole (Can you see intuitively as to why the expectation is so low!?). □

We used a very simple (yet very useful) fact in the above solution:

Fact 6. *Let X be an indicator random variable of an event E of a probability space (Ω, P). Then*

$$\mathbb{E}[X] = P(E).$$

Proof.

$$\mathbb{E}[X] = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = P(X = 1) = P(E).$$

□

[2] Another way to see this is to represent all the outcomes using permutations of 1 to 100, using the uniform distribution on the sample space, and then realizing that the event of R_i going back to H_i is the set of all permutations where i appears in position i, and there are 99! such permutations, and thus the probability of R_i going to H_i is $\frac{99!}{100!} = \frac{1}{100}$.

Lecture 12: Random variables III: conditional expectation, law of total expectation, probability mass functions, some common distributions: Bernoulli, geometric, binomial.

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 19 of the Math for CS book.

1 Linearity of expectation

We will prove the linearity of expectation now:

Theorem 1 (Linearity of expectation)**.** *If $X_1, \ldots, X_n$ are random variables defined on a probability space (Ω, P), $a_1, \ldots, a_n$ and b are real numbers, and $X = (\sum_{i=1}^n a_i X_i) + b$, we have that*

$$\mathbb{E}[X] = \mathbb{E}[(\sum_{i=1}^n a_i X_i) + b] = (\sum_{i=1}^n a_i \mathbb{E}[X_i]) + b.$$

We will prove this in two parts. We will first show the additivity of expectation for two variables (it's easy to make the proof work for more than two variables), and then we will look at how to deal with expectations of random variables with multiplicative and additive constants.

Lemma 2. *Let X and Y be random variables defined on (Ω, P). Then*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Proof. We will use the first definition of expectation. Let $Z = X + Y$. Recall that X, Y, Z are all just functions of the from $\Omega \to \mathbb{R}$, and so we have that for every $\omega \in \Omega$, $Z(\omega) = X(\omega) + Y(\omega)$. Using the definition of $\mathbb{E}[Z]$,

$$\mathbb{E}[Z] = \sum_{\omega \in \Omega} Z(\omega) P(\{\omega\}) = \sum_{\omega \in \Omega} (X(\omega) + Y(\omega)) P(\{\omega\}).$$

We now expand the product to get

$$\mathbb{E}[Z] = \sum_{\omega \in \Omega} (X(\omega)P(\{\omega\}) + Y(\omega)P(\{\omega\})) = \sum_{\omega \in \Omega} X(\omega)P(\{\omega\}) + \sum_{\omega \in \Omega} Y(\omega)P(\{\omega\})$$
$$= \mathbb{E}[X] + \mathbb{E}[Y],$$

where there last equation follows from the definition of $\mathbb{E}[X]$ and $\mathbb{E}[Y]$. □

Lemma 3. *Let X be a random variable defined on a probability space (Ω, P). Let $c, d \in \mathbb{R}$ be constants. Define the random variable $Y = cX + d$. Then*

$$\mathbb{E}[Y] = \mathbb{E}[cX] = c\mathbb{E}[X] + d.$$

Proof. We will use the first definition of expectation:

$$\mathbb{E}[Y] = \sum_{\omega \in \Omega} Y(\omega) P(\{\omega\}).$$

Recall that Y and X are basically functions from Ω to $\mathbb{R}$, and since $Y = cX + d$, for every $\omega \in \Omega$, $Y(\omega) = cX(\omega) + d$. This means that

$$\mathbb{E}[Y] = \sum_{\omega \in \Omega} (cX(\omega) + d) P(\{\omega\}).$$

We can use distributivity to rewrite this as:

$$\mathbb{E}[Y] = \sum_{\omega \in \Omega} (cX(\omega)P(\{\omega\}) + dP(\{\omega\})) = \sum_{\omega \in \Omega} cX(\omega)P(\{\omega\}) + \sum_{\omega \in \Omega} dP(\{\omega\}).$$

Since c and d are constants we can pull them out of the sums:

$$c\left(\sum_{\omega \in \Omega} X(\omega)P(\{\omega\})\right) + d\left(\sum_{\omega \in \Omega} P(\{\omega\})\right) = c\mathbb{E}[X] + d.$$

Here we used the fact that $\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)P(\{\omega\})$ and $\sum_{\omega \in \Omega} P(\{\omega\}) = 1$. □

It's not hard to combine the two lemmas to prove the theorem of linearity of expectation stated above.

Question . Suppose I roll a fair dice 10 times. After each throw, I observe the number rolled, multiply it by 5, and note it down. At the end of all the throws, I add up the numbers I noted down and add 6 to the whole sum. I call this sum X. What's $\mathbb{E}[X]$?

Proof. Let $X_1, \ldots, X_{10}$ denote the random variables that are equal to the numbers rolled in the various throws. Then, we have that

$$X = 5X_1 + 5X_2 + \ldots + 5X_{10} + 6 = 5(\sum_{i=1}^{10} X_i) + 6.$$

Thus, using linearity of expectation we have

$$\mathbb{E}[X] = \mathbb{E}[5(\sum_{i=1}^{10} X_i) + 6] = \mathbb{E}[5(\sum_{i=1}^{10} X_i)] + 6 = 5(\sum_{i=1}^{10} \mathbb{E}[X_i]) + 6.$$

We know that $\mathbb{E}[X_i] = 3.5$ for every i, and so $\mathbb{E}[X] = 181$. □

1.1 Infinite sample spaces

In many cases, the sample space Ω can be countably infinite[1], and/or the random variables can have a countably infinite range. While we wouldn't be studying this case in its full generality in

[1] It can be uncountably infinite but that is beyond the scope of this course.

this course, we will see certain instances of experiments which warrant the use of infinite sample spaces or random variables. The examples we will see are all “nice” cases and so whatever we have learnt about probability and random variables for the finite case will apply to them without any change. The only difference is that sometimes when computing the expectation of such countably infinite random variables, we will have to use infinite series to solve them.

Let’s see an example:

Question . Suppose we keep tossing a fair coin till we see heads. Let X be the total number of tosses till we see a heads. What is $\mathbb{E}[X]$?

Proof. Clearly, Ω in this case is countably infinite: the outcomes are H, TH, TTH,......., and we have that $P(H) = 1/2$, $P(TH) = 1/4$, $P(TTH) = 1/8$, ...

Let X_1 be the random variable that’s 1 if the first flip is a tails and it’s 0 otherwise. Let X_2 be the random variable which is 1 if the first and second flips results in tails, and 0 otherwise. Similarly, we can define X_i to be 1 if the first i coin tosses are tails and 0 otherwise. We will define X_i for every natural number i and so we have a countably infinite number of variables. Then, if X is the number of times we flip before we see a heads:

$$X = 1 + X_1 + X_2 + \ldots + (\text{to } \infty).$$

You can convince yourself that the above equation is indeed true: just try to analyze the different cases; the case when you see a heads in the first toss, the case when you see a heads in the second toss, and so on. Then, using linearity of expectation (Let’s assume it works for infinitely many random variables without getting into the details):

$$\mathbb{E}[X] = 1 + \sum_{i=1}^{\infty} \mathbb{E}[X_i].$$

What is $\mathbb{E}[X_i]$? Since X_i is an indicator random variable (for what event?why?), we know that

$$\mathbb{E}[X_i] = P(X_i = 1) = P(\text{the first } i \text{ coin tosses result in tails}) = \left(\frac{1}{2}\right)^i.$$

Here the last equation follows from the fact that all the coin tosses are independent. So we have

$$\mathbb{E}[X] = 1 + \sum_{i=1}^{\infty} \frac{1}{2^i} = \sum_{i=0}^{\infty} \frac{1}{2^i}.$$

Recall from calculus that for $0 < a < 1$:

$$\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}.$$

In our case $a = 1/2$, and so we get

$$\mathbb{E}[X] = 2.$$

□

2 Conditional expectation

Let X be a random variable defined on a probability space (Ω, P), and let A be an event of the probability space. Then, we can define the expectation of X *given* that the event A has happened. A natural way to do this is to look at the expression for expectation of X and replace probability with the conditional probability. In particular, recall the second definition of expectation:

$$\mathbb{E}[X] = \sum_{a \in Range(X)} aP(X = a).$$

To define the *conditional expectation of* X, i.e. the expectation conditioned on the event A, we simply replace $P(X = a)$ by $P(X = a|A)$, to get

$$\mathbb{E}[X|A] = \sum_{a \in Range(X)} aP(X = a|A).$$

Question . Suppose we keep throwing a dice till we see a 6. Let Y be the total number of throws needed, and let X be the total number of ones observed. Compute $\mathbb{E}[X|Y = 10]$.

Proof. Since we have to find the expectation of X given that $Y = 10$, we can only focus on the first 10 throws. In fact, since we know that the 10^{th} dice must be a 6, we can just focus on the first 9 throws. Let $X_1, \ldots, X_9$ be random variables such that X_i is the indicator random variable for the event that the i^{th} throw results in a 1. Then, we have that

$$\mathbb{E}[X|Y = 10] = \mathbb{E}[\sum_{i=1} X_i|Y = 10].$$

This is because once we are given the information that $Y = 10$, X can be written as $X_1+X_2+\ldots+X_9$ (why?).

Also to be noted is the fact that linearity of expectation holds even in the case of conditional expectation (the same proof as before will work. One simply has to replace probability by conditional probability). Thus,

$$\mathbb{E}[X|Y = 10] = \mathbb{E}[\sum_{i=1}^{9} X_i|Y = 10] = \sum_{i=1}^{9} \mathbb{E}[X_i|Y = 10].$$

We now need to compute $\mathbb{E}[X_i|Y = 10]$. Let's just use the definition of conditional expectation to do so:

$$\mathbb{E}[X_i|Y = 10] = \sum_{a \in Range(X_i)} aP(X_i = a|Y = 10) = 1 \cdot P(X_i = 1|Y = 10) + 0 \cdot P(X_i = 0|Y = 10).$$

$$= P(X_i = 1|Y = 10) = \frac{P(X_i = 1 \cap Y = 10)}{P(Y = 10)}.$$

□

It's not hard to compute that $P(Y = 10) = (5/6)^9(1/6)$, and $P(X_i = 1 \cap Y = 10) = (5/6)^8(1/6)^2$, and this gives

$$\mathbb{E}[X_i|Y = 10] = P(X_i = 1|Y = 10) = \frac{(5/6)^8(1/6)^2}{(5/6)^9(1/6)} = \frac{1}{5}.$$

This means that

$$\mathbb{E}[X|Y = 10] = \sum_{i=1}^{9} \frac{1}{5} = \frac{9}{5}.$$

3 Law of total expectation

The way we can compute a probability of an event B by computing the probability of B happening conditioned on disjoint events $A_1, \ldots, A_n$ that form a parition of Ω i.e. compute $P(B|A_1), P(B|A_2), \ldots$, and then combining all these cases using the law of total probability, i.e. $P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$, we can also do the same for expectation.

Theorem 4 (Law of total expectation)**.** *Consider a probability space (Ω, P) and let $A_1, \ldots A_n$ be disjoint nonempty events in it that form a partition of Ω. Then,*

$$\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i)\mathbb{E}[X|A_i].$$

Proof. The proof follows from the law of total probability and the second definition of expectation:

$$\mathbb{E}[X] = \sum_{k \in Range(X)} kP(X = k).$$

Using law of total probability, we can write

$$P(X = k) = \sum_{i=1}^{n} P(A_i)P(X = k|A_i).$$

Substituting this back into the first equation, we get

$$\mathbb{E}[X] = \sum_{k \in Range(X)} k \left(\sum_{i=1}^{n} P(A_i)P(X = k|A_i) \right) = \sum_{k \in Range(X)} \sum_{i=1}^{n} kP(A_i)P(X = k|A_i).$$

We can now switch the sums to get

$$\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i) \left(\sum_{k \in Range(X)} kP(X = k|A_i) \right).$$

But the expression inside the paranthesis is just $\mathbb{E}[X|A_i]$, and so we get

$$\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i)\mathbb{E}[X|A_i].$$

□

Question . Suppose we keep tossing a biased coin (the coin lands heads with probability p and tails with probability $1-p$) till we see a heads. On average, how many times will we have to toss the coin?

Proof. Let X be the number of times the coin is tossed till a heads is observed. Let A be the event that we observe a heads in the very first coin toss. Then using the law of total expectation:

$$\mathbb{E}[X] = P(A)\mathbb{E}[X|A] + P(A^c)\mathbb{E}[X|A^c].$$

Here $P(A) = p$, $P(A^c) = 1-p$, $\mathbb{E}[X|A] = 1$ (why?). How about $\mathbb{E}[X|A^c]$? Notice that if the first coin toss is not heads, then *starting from the second coin toss*, we will need the same number of coin tosses on average to see a heads as we would if starting from the first coin toss (Think about this - it's important!). Basically, if our first toss results in tails, it's almost like forgetting that this happened and starting all over again, and so

$$\mathbb{E}[X|A^c] = 1 + \mathbb{E}[X].$$

The one is added to the RHS to account for the first coin toss that resulted in tails. Plugging in these values in the expression for $\mathbb{E}[X]$, we get

$$\mathbb{E}[X] = p \cdot 1 + (1-p)(1 + \mathbb{E}[X]).$$

Let $z = \mathbb{E}[X]$. We now have a linear equation in the variable z, and we can solve for it:

$$z = p + (1-p)(1+z) \implies z = p + 1 - p + z(1-p)$$
$$\implies pz = 1 \implies z = \frac{1}{p}$$

Thus, $\mathbb{E}[X] = \frac{1}{p}$. □

4 Probability mass function

Before we begin let's set some things straight. Up till now, when we have been talking about probability in the context of probability spaces (Ω, P), we have been calling them "probability distributions". While it's okay use to this term (and some people do), (I think) the more widely used term is "probability measure". I did not want to introduce this in the beginning to avoid confusion/intimidation.

Formally, with an experiment we associate a sample space Ω and then we define a probability measure P on Ω (what we have been calling a probability distribution so far). A probability measure is basically a function $P : 2^\Omega \to \mathbb{R}$, a function that assigns a "probability" with every event.

Having set the record straight, we can now define what we mean by a probability mass function.

Definition 5 (Probability mass function of a random variable)**.** *Let X be a random variable defined on a probability space (Ω, P) (remember P is called a probability measure on Ω)). The probability mass function f_X : Range$(X) \to \mathbb{R}$ is a function that assigns to every value in the range of X the probability of X being equal to that value, i.e. for every $a \in$ Range(X),*

$$f_X(a) = P(X = a).$$

So basically, $f_X(a)$ is just a fancy way of saying $P(X = a)$, for all we care[2].

In some sense, the probability mass function is like providing a table of probabilities for all the values in the range of X. For example, consider the experiment of tossing a coin three times, and let X be the number of heads observed. Then, $Range(X) = \{0, 1, 2, 3\}$, and we have

$$f_X(0) = P(X = 0) = \frac{1}{8}$$

$$f_X(1) = P(X = 1) = \frac{3}{8}$$

$$f_X(2) = P(X = 2) = \frac{3}{8}$$

$$f_X(3) = P(X = 3) = \frac{1}{8}$$

To see interesting properties of the probability mass function, you basically have to just look for interesting properties of $P(X = a)$. We already did this lecture 11.

5 Probability distributions

A random variable X defined on some probability space (Ω, P) along with its probability mass function f_X is called a *probability distribution* (Remember, we decided to call P a probability measure from now on). There are some common probability distributions that arise quite often when modelling experiments and solving problems, and so it's good to be aware of them.

5.1 Bernoulli distribution

A Bernoulli distribution with parameter p is a random variable X with range $\{0, 1\}$ defined on a probability space (Ω, P), along with its probability mass function f_X defined as

$$f_X(1) = P(X = 1) = p$$

$$f_X(0) = P(X = 0) = 1 - p.$$

We typically call X a *Bernoulli random variable*, or say that X is distributed according to the Bernoulli distribution. Common situations where Bernoulli random variables/distributions arise are coin tosses and indicator random variables.

The expected value of a Bernoulli random variable is

$$\mathbb{E}[X] = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = P(X = 1) = f_X(1) = p.$$

[2] In the case of continous random functions, the counterpart of the probability mass function stops being just a fancy alternative to $P(X = a)$, as you will see if you take a course on probability at some point.

5.2 Geometric distribution

A geometric distribution with parameter p is a random variable X with range $\{1, 2, 3, 4, \ldots\}$ defined on an infinite probability space (Ω, P), along with its probability mass function f_X defined as

$$f_X(k) = P(X = k) = (1-p)^{k-1}p$$

for $k \in \mathit{Range}(X) = \{1, 2, 3, 4, \ldots, \}$. A good way to think of geometric distributions is the experiment of tossing a biased coin (which has probability of heads being equal to p, and tails $1-p$) till we observe heads, and thinking of X as the number of tosses. We saw in section 3 (see the example there), that if X is a geometric random variable with parameter p, then $\mathbb{E}[X] = \frac{1}{p}$.

5.3 Binomial distribution

A binomial distribution with parameters n and p is a random variable X with range $\{0, 1, \ldots, n\}$ defined on a probability space (Ω, P), along with its probability mass function f_X defined as follows:

$$f_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k \in \mathit{Range}(X) = \{0, 1, \ldots, n\}$. A good way to think of the binomial distribution is to think of tossing a biased coin (which turns up heads with probability p and tails with $1-p$) n times and thinking of X as the number of heads observed. Suppose we define indicator (and so Bernoulli) random variables $X_1, \ldots, X_n$ such that X_i is 1 if and only if the i$^{\text{th}}$ coin toss results in heads, then clearly

$$X = X_1 + \ldots + X_n.$$

Thus, a binomial random variable can be thought of as the sum of n independent Bernoulli random variables, each having parameter p. This also means that

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

using linearity of expectation, and since each $\mathbb{E}[X_i] = p$ (since they are all Bernoulli), we get that for a binomial distribution the expected value is np, i.e. $\mathbb{E}[X] = np$.

To see why $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, let's see what the sample space and probability measure are in this case. Let's represent the outcomes using binary strings of length n, where a 1 in the i$^{\text{th}}$ position indicates that the i$^{\text{th}}$ coin toss results in heads, a 0 indicates that the i$^{\text{th}}$ coin toss results in tails. Thus, Ω is the set of all binary strings of length n.

How is the probability measure P on Ω defined? For an outcome (binary string) $\omega \in \Omega$, if ω has k ones and $n-k$ zeros, since the probability of heads is p and of tails is $1-p$, and all the tosses are independent, we have that

$$P(\{\omega\}) = p^k (1-p)^{n-k}.$$

(why is the above true? Try to use the chain rule for independent events if you can't see why!) The event $[X = k]$, i.e. seeing exactly k heads, consists of all binary strings (outcomes) with k ones and $n-k$ zeros. Since there are $\binom{n}{k}$ such outcomes, and they are all disjoint (outcomes are basically atomic events, i.e. events that are singleton sets, and so are always disjoint from each other!) and each outcome has probability $p^k(1-p)^{n-k}$, using the sum rule, we can observe that

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

Lecture 13 (Part 1): Random variables IV: expectation of products of random variables, covariance

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 19 of the Math for CS book.

Before we study expectation of products, let's look at another distribution, called the uniform distribution (this is different from the uniform *measure* P on a sample space Ω which assigns probability $\frac{1}{|\Omega|}$ to each outome, i.e. all outcomes are equally likely).

1 Uniform distribution

A uniform distribution is a random variable X with range $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}$ (defined on a probability space (Ω, P)) with probability mass function f_X defined as follows: for every $a \in Range(X)$,

$$f_X(a) = P(X = a) = \frac{1}{n}.$$

This just means that X is equally likely to take any of the n values in its range. This also means that the expected value of X is

$$\mathbb{E}[X] = \sum_{a \in Range(X)} aP(X = a) = \sum_{i=1}^{n} a_i P(X = a_i) = \sum_{i=1}^{n} a_i \frac{1}{n} = \frac{\sum_{i=1}^{n} a_i}{n}.$$

2 Expectation of products of random variables

Suppose X, Y are random variables defined on (Ω, P). What is $\mathbb{E}[XY]$? Turns out that if X and Y are independent, then there is an easy way to compute this:

Theorem 1. *Let X and Y be independent random variables then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

Proof. We will use the law of total expectation to rewrite $\mathbb{E}[XY]$:

$$\mathbb{E}[XY] = \sum_{b \in Range(Y)} P(Y = b)\mathbb{E}[XY | Y = b],$$

which works because the events $[Y = b]$ for $b \in Range(Y)$ partition Ω. Also, note that given $Y = b$, the random variable XY is just bX. So we get,

$$\mathbb{E}[XY] = \sum_{b \in Range(Y)} P(Y = b)\mathbb{E}[bX | Y = b] = \sum_{b \in Range(Y)} P(Y = b) b \mathbb{E}[X | Y = b].$$

Here the last equation uses the linearity of conditional expectation. Let's substitute the expression for $\mathbb{E}[X|Y=b]$ in the above expression:

$$\mathbb{E}[XY] = \sum_{b \in Range(Y)} P(Y=b)\mathbb{E}[bX|Y=b] = \sum_{b \in Range(Y)} P(Y=b)b\left(\sum_{a \in Range(X)} aP(X=a|Y=b)\right).$$

Since X and Y are independent, we know that $P(X=a|Y=b) = P(X=a)$, and so

$$\mathbb{E}[XY] = \sum_{b \in Range(Y)} P(Y=b)b\left(\sum_{a \in Range(X)} aP(X=a|Y=b)\right) = \sum_{b \in Range(Y)} P(Y=b)b\left(\sum_{a \in Range(X)} aP(X=a)\right)$$

$$= \sum_{b \in Range(Y)} P(Y=b)b\left(\mathbb{E}[X]\right).$$

Here the last equation follows from the definition of $\mathbb{E}[X]$. Since $\mathbb{E}[X]$ is a constant that does not depend on b we can pull it out, and so

$$\mathbb{E}[XY] = \mathbb{E}[X]\left(\sum_{b \in Range(Y)} bP(Y=b)\right) = \mathbb{E}[X]\mathbb{E}[Y].$$

□

More generally, for n mutually independent random variables, we have the following

Theorem 2. *If $X_1, \ldots, X_n$ are mutually independent random variables, then*

$$\mathbb{E}[X_1 X_2 \ldots X_n] = \mathbb{E}[X_1]\mathbb{E}[X_2]\ldots\mathbb{E}[X_n].$$

Question . Let $p_1, \ldots, p_k$ be mutually independent random variables such that each p_i is a uniformly random prime number between 1 and 20 (1 is not a prime). If $X = p_1 p_2 \ldots p_k$, what is $\mathbb{E}[X]$?

Proof. Since all n random variables are mutually independent, we know that $\mathbb{E}[X] = \mathbb{E}[p_1]\mathbb{E}[p_2]\ldots\mathbb{E}[p_k]$. It's easy to see that for each $i \in \{1, \ldots, k\}$, p_i is a uniform random variable with range $\{2, 3, 5, 7, 11, 13, 17, 19\}$, and so

$$\mathbb{E}[p_i] = \frac{2+3+5+7+11+13+17+19}{8} = 9.625,$$

and so $\mathbb{E}[X] = (9.625)^k$. □

Often we deal with n variables that are mutually independent and each variable has the same probability distribution. Such variables are called *Independent Identically Distributed (i.i.d.) random variables.*

3 Covariance of random variables

Given two random variables X, Y defined on a probability space, the *covariance* between them is defined as follows:

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

We will first derive an equivalent expression for covariance that will be easier to interpret:

Lemma 3.

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Proof. Let's write $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. It's important to recall that μ_X and μ_Y are fixed values and not random variables. Then,

$$Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] = \mathbb{E}[XY] - \mu_X \mathbb{E}[Y] - \mu_Y \mathbb{E}[X] + \mu_X \mu_Y.$$

Here the last step follows from linearity of expectation. Putting back $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$, we get

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

□

Clearly, if X and Y are independent, then based on what we saw in the previous section and based on the alternate expression for covariance that we derived, $Cov(X, Y) = 0$. Is the other way round true? That is, if $Cov(X, Y) = 0$, then X and Y are independent? This is not true. Consider X with range $\{-1, 0, 1\}$ and probability mass function f_X defined as follows:

$$f_X(-1) = P(X = -1) = \frac{1}{4}$$

$$f_X(0) = P(X = 0) = \frac{1}{2}$$

$$f_X(1) = P(X = 1) = \frac{1}{4},$$

and let $Y = X^2$. Then, $\mathbb{E}[X] = 0$ (why?), and so $\mathbb{E}[X]\mathbb{E}[Y] = 0$. How about $\mathbb{E}[XY]$? Note that

$$\mathbb{E}[XY] = \mathbb{E}[X^3] = (-1)\frac{1}{4} + (0)\frac{1}{2} + (1)\frac{1}{4} = 0.$$

But are X and Y are independent? No, because Y is literally defined as X^2 and so, for example, $P(Y = 1 | X = 0) = 0 \neq P(Y = 1)$.

Nevertheless, covariance can be thought of as a measure of how dependent two random variables are on each other. The farther $Cov(X, Y)$ is from 0, the more "dependent" X and Y are on each other. In fact, it can be shown that if the covariance is "far enough" from 0 (i.e., it is a large enough positive value, or a small enough negative value) then Y can be written as a linear function of X, i.e. $Y = aX + b$ for some constants a and b in $\mathbb{R}$. Conversely, if $Cov(X, Y) = 0$ then all that means is that Y cannot be written as a linear function of X, i.e. Y does not have a linear dependence on X (but as we saw, Y can still be dependent on X in a way other than linear, e.g., $Y = X^2$).

Lecture 13 (Part 2): Deviation from mean: Markov's inequality, variance and its properties, Chebyshev's inequality

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 20 of the Math for CS book.

So far we have learned how to model random experiments/phenomenon mathematically using probability theory: define what the outcomes are and the sample space Ω, define an appropriate probability measure on Ω, identify relevant events, define useful random variables, and compute expected value of random variables. But simply computing the expected value of a random variable does provide any guarantees about the "typical" values that a random variable takes. We want to be able to make predictions about the kind of values the random variable will take, and we want those predictions to hold with high probability.

Formally speaking, suppose we have a random variable X defined on some probability space (Ω, P). We want to be able make statements of the form:

"X will be in the range $[a, b]$ with probability at least α",

or

$$P(a \leq X \leq b) \geq \alpha.$$

Ideally, we would want a and b to be as close to each other as possible, i.e. the interval $[a, b]$ to be as small as possible, and simultaneously, the probability α to be as high as possible (for example, $\alpha \geq 0.99$).

For example, suppose you design a randomized algorithm for a problem whose running time T is a random variable, i.e. it depends on some random choices the algorithm makes when it's run on any given input. Also suppose that $\mathbb{E}[T]$, the expected running time of the algorithm, is small, and so you want to be able to make a statement of the form "with high probability, the running time T of the algorithm is "close" to the expected running time $\mathbb{E}[T]$." because that shows that most of the time your algorithm runs very fast (you will encounter such situations when you study and design algorithms in CS 344).

In this lecture, we will see how one can provide guarantees on the typical value that random variables take. In particular, we will show that for many distributions it's highly unlikely that a random variable takes a value that's "too far" from it's expected value, i.e. it's highly unlikely that the random variable *deviates* too much from it's expected value!

1 Markov's inequality

Suppose X is a random variable defined on a probability space (Ω, P) such that X always takes nonnegative values, i.e. for every $a \in \mathit{Range}(X)$, $a \geq 0$. We call X a nonnegative random variable.

Theorem 1 (Markov's inequality)**.** *Let X be a nonnegative random variable with expectation $\mathbb{E}[X]$, and let $a \geq 0$ be a constant. Then*

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Equivalently, the probability that X becomes $k\mathbb{E}[X]$ (for some $k \geq 1$) is at most $\frac{1}{k}$:

$$P(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}.$$

Proof. We will prove the first statement of the theorem. The second statement follows from the first by setting $a = k\mathbb{E}[X]$. Using the second definition of $\mathbb{E}[X]$:

$$\mathbb{E}[X] = \sum_{i \geq 0} iP(X = b) = \sum_{i=0}^{a-1} iP(X = i) + \sum_{j=a}^{\infty} jP(X = j).$$

Clearly, $\sum_{i=0}^{a-1} iP(X = i) \geq 0$, and so

$$\mathbb{E}[X] \geq \sum_{j=a}^{\infty} jP(X = j) \geq \sum_{j=a}^{\infty} aP(X = j) = a \left(\sum_{j=a}^{\infty} P(X = j) \right) = aP(X \geq a),$$

where the last step follows from the fact that $P(X \geq a) = \sum_{j=a}^{\infty} P(X = j)$. Thus, this implies that

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

□

Question . Consider the experiment of tossing a biased coin (probability of heads is 1/3) 100 times. What is the probability that the number of heads observed is less than 70?

Proof. This experiment can be modelled using a binomial random variable X with $n = 100$ and $p = 1/3$ (why?), and so $\mathbb{E}[X] = np = \frac{100}{3} \approx 33.33$. Using Markov's inequality,

$$P(X \geq 70) \leq \frac{\mathbb{E}[X]}{70} \leq \frac{33.33}{70} \leq 0.48.$$

So, with probability at least 0.52, the number of heads is at most 70. □

2 Variance

A useful quantity to understand that helps prove somewhat better guarantees for random variables is the *average deviation of a random variable from its mean*[1].

Of course, if we define deviation from the expected value naively, i.e. as $X - \mathbb{E}[X]$, then the average deviation will be zero: $\mathbb{E}[(X - \mathbb{E}[X])] = \mathbb{E}[X] - \mathbb{E}[X] = 0$. The reason that this naive definition becomes zero is because it doesn't consider the magnitude of deviation and instead considers the

[1] The expected value is also called the mean of a random variable

signed value of deviation. For example, if there is a random variable X with range $\{-1000, 1000\}$ such that $P(X = -1000) = P(X = 1000) = 1/2$, then $\mathbb{E}[X] = 0$. Now if we define deviation as $X - \mathbb{E}[X]$, then when $X = -1000$, the deviation becomes -1000, and when X is 1000, the deviation becomes 1000, and so the deviations cancel each other out when we average them.

Intuitively, what we would want as the average deviation in this case would be $\frac{1000+1000}{2} = 1000$. This suggests we define deviation as $|X - \mathbb{E}[X]|$, where $|\cdot|$ is the absolute value function. In this case, the average/expected deviation would be $\mathbb{E}[|X - \mathbb{E}[X]|]$. While this is a sensible quantity to study, it turns out that it's not very friendly to mathematical manipulation and analysis. Instead what's preferred is to consider the square of deviation, i.e. $(X - \mathbb{E}[X])^2$, and then compute its average/expected value. This is called the *variance* of X.

Definition 2 (Variance). *The variance of a random variable X, denoted by $Var(X)$, with expected value $\mathbb{E}[X]$ is the average squared deviation of X from $\mathbb{E}[X]$:*

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

There is an alternate expression that we can derive for the variance that makes computing variance easy in many cases:

Lemma 3. *For a random variable X with mean $\mathbb{E}[X]$,*

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

Proof. Notice from the definition of variance that

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])] = Cov(X, X),$$

using the second definition of covariance. Using the first definition of covariance we know that

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Thus, $Cov(X, X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, and so

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

□

Notice that variance is the average of the square of the deviation of a random variable from its mean, and so to bring it down to the same scale as the mean, it's often useful to consider the square root of the variance. The square root of the variance is called the *standard deviation*, and is denoted be $\sigma(X)$ for a random variable X, i.e.

$$\sigma(X) = \sqrt{Var(X)}.$$

To compute the variance using the second expression, we need to know the value of $\mathbb{E}[X^2]$ for a random variable. But this is easy to compute (do you see why? Think of X^2 as a new random variable!):

$$\mathbb{E}[X^2] = \sum_{a \in Range(X)} a^2 P(X = a).$$

The quantity $\mathbb{E}[X^2]$ is called the *second moment* of X. In general, one can study $\mathbb{E}[X^k]$ for $k \geq 1$, and this quantity is called the k^{th} *moment* of X, and can be computed in a similar way:

$$\mathbb{E}[X^k] = \sum_{a \in Range(X)} a^k P(X = a).$$

Question . Let X be the number rolled by a fair dice. Find $Var(X)$.

Proof. Recall that $\mathbb{E}[X] = 3.5$. We will use the formula $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. We already know the second term, so we only need to compute $\mathbb{E}[X^2]$. We will use the formula for the second moment to do this (see above):

$$\mathbb{E}[X^2] = \sum_{a \in Range(X)} a^2 P(X = a) = \sum_{i=1}^{6} i^2 P(X = i) = \sum_{i=1}^{6} i^2 \frac{1}{6} = \frac{\sum_{i=1}^{6} i^2}{6}.$$

The formula for the sum of the squares of numbers from 1 to n is

$$\frac{n(n+1)(2n+1)}{6}.$$

Thus,

$$\mathbb{E}[X^2] = \frac{\frac{6(6+1)(12+1)}{6}}{6} = \frac{91}{6}.$$

Thus,

$$Var(X) = \frac{91}{6} - (3.5)^2 \approx 2.91.$$

□

2.1 Properties of variance

Here are some properties of variance:

1. If $X : \Omega \to \mathbb{R}$ is a constant random variable, i.e. $\forall \omega \in \Omega$, $X(\omega) = c$, then $Var(X) = 0$ (check this using the definition of variance). In other words, for any constant $c \in \mathbb{R}$, $Var(c) = 0$.

2. Let $a \in \mathbb{R}$ be a constant and X be a random variable, then $Var(aX) = a^2 Var(X)$. This follows because $Var(aX) = \mathbb{E}[(aX)^2] - (\mathbb{E}[aX])^2 = \mathbb{E}[a^2X^2] - (a\mathbb{E}[X])^2 = a^2\mathbb{E}[X^2] - a^2(\mathbb{E}[X])^2 = a^2(\mathbb{E}[X^2] - (\mathbb{E}[X])^2) = a^2 Var(X)$.

3. This one is very useful: if $X_1, \ldots, X_n$ are pairwise independent random variables, then

$$Var(X_1 + \ldots + X_n) = Var(X_1) + Var(X_2) + \ldots + Var(X_n).$$

   Note that we only need $X_1, \ldots, X_n$ to be pairwise independent[2] and not mutually independent (the latter implies the former). To get an idea as to why this is true, let's look at the case when $n = 2$ (the general case then follows from induction). Let's look at $Var(X + Y)$ for

[2] It's important to contrast this with linearity of expectation where $X_1, \ldots, X_n$ did not need any independence whatsoever.

two independent random variables X and Y. Using the definition of variance and linearity of expectation,

$$Var(X+Y) = \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 = \mathbb{E}[X^2+Y^2+2XY] - (\mathbb{E}[X]+\mathbb{E}[Y])^2$$
$$= \mathbb{E}[X^2]+\mathbb{E}[Y^2]+2\mathbb{E}[XY] - \big((\mathbb{E}[X])^2+(\mathbb{E}[Y])^2+2\mathbb{E}[X]\mathbb{E}[Y]\big)$$

Rearranging terms, we get

$$Var(X+Y) = (\mathbb{E}[X^2]-(\mathbb{E}[X])^2) + (\mathbb{E}[Y^2]-(\mathbb{E}[Y])^2) + 2(\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y])$$
$$\implies Var(X+Y) = Var(X)+Var(Y)+2Cov(X,Y).$$

Since X and Y are independent, $Cov(X,Y)=0$, and so

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

4. In general, we have that, for constants $a_1,\dots,a_n,b\in\mathbb{R}$, and pairwise independent random variables $X_1,\dots,X_n$,

$$Var(\sum_{i=1}^{n} a_iX_i+b) = \sum_{i=1}^{n} a_i^2 Var(X_i).$$

Question . A fair dice is rolled 100 times. Let X be the sum of the numbers observed. What is $Var(X)$?

Proof. Let X_i be the number observed in the ith roll. Then $X=\sum_{i=1}^{100}X_i$. Using the properties of variance,

$$Var(X) = \sum_{i=1}^{100} Var(X_i).$$

We computed the variance of a single dice roll at the end of the previous subsection: $Var(X_i) = \frac{91}{6}-\left(\frac{3}{2}\right)^2$, and so

$$Var(X) = 100\times\left(\frac{91}{6}-\left(\frac{3}{2}\right)^2\right)\approx 291.67$$

□

2.2 Variance of some common distributions

We will now revisit some common distributions/random variables to see what their respective variance are:

2.2.1 Bernoulli

Recall that a Bernoulli random variable X is 1 with probability p and 0 with probability $1-p$, and $\mathbb{E}[X]=p$. Let's compute $Var(X)$:

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

Note that $\mathbb{E}[X^2] = 1\cdot P(X=1)+0\cdot P(X=0)=p$, and so

$$Var(X) = p-p^2 = p(1-p).$$

2.2.2 Binomial

Recall that a Binomail random variable is basically a sum of n independent Bernoulli random variables with parameter p, i.e. $X = X_1 + \ldots + X_n$, where each X_i is Bernoulli with parameter p, i.e. $P(X = 1) = p$ and $P(X = 0) = (1 - p)$. Using the properties of variance, we know that

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

Since each X_i is Bernoulli with parameter p, we have that $Var(X_i) = p(1 - p)$, and so

$$Var(X) = np(1 - p).$$

Thus, the variance of a binomial random variable with parameter p is $np(1 - p)$.

2.2.3 Geometric

Recall that a geometric random variable with parameter p is the basically the total number of coin flips in an experiment in which we keep tossing a biased coin (i.e., the probabilty of heads is p) till we see a heads. Using the law of total expectation, we proved that if X is a geometric random variable with parameter p, then $\mathbb{E}[X] = \frac{1}{p}$. Using a similar proof, one can show that

$$Var(X) = \frac{1-p}{p^2}.$$

To see how, we use the fact that $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[X^2] - \frac{1}{p^2}$. So we basically have to compute $\mathbb{E}[X^2]$ for a geometric random variable with parameter p. Let A be the event that the first coin toss is heads. Then using the the law of total expectation:

$$\mathbb{E}[X^2] = P(A)\mathbb{E}[X^2|A] + P(A^c)\mathbb{E}[X^2|A^c].$$

It's easy to see that $\mathbb{E}[X^2|A] = 1$ (why?). Also, $P(A) = p$ and $P(A^c) = 1-p$. How about $\mathbb{E}[X^2|A^c]$? Once we know that the first coin toss is tails we basically have to start over, and if Y is the total number of tosses, *starting with the second coin toss* (i.e., not counting/including the first coin toss), needed to see a heads, then

$$\mathbb{E}[X^2|A^c] = \mathbb{E}[(1 + Y)^2].$$

The key observation is that Y is also a geometric random variable with parameter p, and in fact, has the same probability mass function as X, and so $\mathbb{E}[(1+Y)^2] = \mathbb{E}[(1+X)^2]$. Putting everything together, we get

$$\mathbb{E}[X^2] = p \cdot 1 + (1 - p)\mathbb{E}[(1 + X)^2] = p + (1 - p)\left(\mathbb{E}[X^2 + 1 + 2X]\right)$$
$$= p + (1 - p)\left(\mathbb{E}[X^2] + 1 + 2\mathbb{E}[X]\right) = p + (1 - p)\mathbb{E}[X^2] + (1 - p) + 2(1 - p)\mathbb{E}[X].$$

Using the fact that $\mathbb{E}[X] = 1/p$, and rearranging terms we get

$$p\mathbb{E}[X^2] = \frac{2-p}{p} \implies \mathbb{E}[X^2] = \frac{2-p}{p^2}.$$

This implies that

$$Var(X) = \frac{2-p}{p^2} - \frac{1}{p} = \frac{1-p}{p^2}.$$

3 Chebyshev's inequality

Chebyshev's inequality gives a much stronger guarantee on how much a random variable can deviate from its expectation if we know the variance of the random variable.

Theorem 4 (Chebyshev's inequality)**.** *Let X be a random variable with variance $Var(X)$ and expectation $\mathbb{E}[X]$. Then*

$$P(|X - \mathbb{E}[X]| \geq a) \leq \frac{Var(X)}{a^2}.$$

Equivalently,

$$P(|X - \mathbb{E}[X]| \geq k \cdot Var(X)) \leq \frac{1}{k^2}.$$

Proof. As in the case of Markov's inequality, we will prove the first statement, because the second statement follows from the first one by setting $a = k\mathbb{E}[X]$. Let Y be equal to the square of the deviation of X from $\mathbb{E}[X]$:

$$Y = (X - \mathbb{E}[X])^2$$

Note that Y is also a random variable (why?), and is in fact a nonnegative random variable (why?), and from the definition of variance, we know that

$$\mathbb{E}[Y] = Var(X).$$

Also, note that

$$P(|X - \mathbb{E}[X]| \geq a) = P((X - \mathbb{E}[X])^2 \geq a^2) = P(Y \geq a^2),$$

where the first equality is true since we can just square both sides of the inequality. Thus, if we can prove an upper bound on $P(Y \geq a^2)$, that will also prove an upper bound on $P(|X - \mathbb{E}[X]| \geq a)$. We can now use Markov's inequality to bound $P(Y \geq a^2)$ (we can use Markov's because Y is a nonnegative random variable):

$$P(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a^2} = \frac{Var(X)}{a^2}.$$

Thus,

$$P(|X - \mathbb{E}[X]| \geq a) = P(Y \geq a^2) \leq \frac{Var(X)}{a^2}.$$

□

If the variance of X is small, then Chebyshev's inequality provides a much stronger guarantee as compared to Markov's inequality:

- With Markov's inequality you can make predictions of the form "X is between 0 and $k\mathbb{E}[X]$" with probability at least $1 - \frac{1}{k}$",
- whereas with Chebyshev's inequality you can make predictions of the form "X is between $\mathbb{E}[X] - k \cdot Var(X)$ and $\mathbb{E}[X] + k \cdot Var(X)$ with probability at least $1 - \frac{1}{k^2}$".

Note that firstly, the probability of the predictions made by Chebyshev's is much better (larger) than that of Markov's: $k \geq 0$ as an integer, and then clearly, $1 - \frac{1}{k} < 1 - \frac{1}{k^2}$. For example, when $k = 10$, $1 - \frac{1}{k} = 0.90$, while $1 - \frac{1}{k^2} = 0.99$!

Secondly, the range of the prediction made by Chebyshev's is smaller than that of Markov's: while Markov's says that X can be anywhere between 0 and $k\mathbb{E}[X]$ (which is a pretty big range!), Chebyshev says that X has to be between $\mathbb{E}[X] - k \cdot Var(X)$ and $\mathbb{E}[X] + k \cdot Var(X)$ which can be pretty small if $Var(X)$ is small.

Let's revisit an earlier problem and try to solve it using Chebyshev's inequality:

Question . Consider the experiment of tossing a biased coin (probability of heads is 1/3) 100 times. What is the probability that the number of heads observed is less than 70?

Proof. The number of heads, say we call it X, is a binomial random variable with parameter p, and so $\mathbb{E}[X] = \frac{100}{3} \approx 33.33$, and $Var(X) = 100 \cdot \frac{1}{3}\left(1 - \frac{1}{3}\right) = \frac{200}{9} \approx 22.22$. We want to upper bound $P(X > 70)$ which is the same as $P(X - \mathbb{E}[X] > 70 - 33.33) = P(X - \mathbb{E}[X] > 36.67) \leq P(|X - \mathbb{E}[X]| > 36.67)$. Using Chebyshev's inequality,

$$P(|X - \mathbb{E}[X]| > 36.67) \leq \frac{Var(X)}{(36.67)^2} = \frac{22.22}{(36.67)^2} \leq 0.0165.$$

This means that the probability the numbers of heads is more than 70 is at most 0.016, and so with probability greater than 0.98, X will be less than 70! This is a much better guarantee than what we got using Markov's inequality. □