Computer Science

Dr. Y. Liow (July 23, 2020)

Contents

Chapter 9

Probability

File: chap-discrete-probability.tex

File: discrete-probability/prob.tex

9.1 Probabilistic or random experiments

9.1.1 Coin toss experiments

If you have a coin, you can toss it and get either head or tail. (It won't land on its side ... unless if you have an extremely *thick* coin ...) If the coin is fair, half the time you will get a head and half the time you will get a tail. Let's call each toss a random experiment and let's call the result of your toss the outcome of the experiment. The two possible outcomes are getting a head and getting tail. I'll create symbols to denote these two outcomes:

Head and Tail

If you're the gambling kind, you know that in fact there's a slightly higher chance that the head falls face down so that there's a slightly higher chance of getting the tail. Let's ignore this fact and just assume that our coin is absolutely fair.

If you don't have a fair coin, you can write a program simulate the tossing of the coin like this one in your favorite programming language (C/C++, Java, Python, what-have-you):

```
import random; random.seed()

HEAD = "HEAD"
TAIL = "TAIL"
S = [HEAD, TAIL] # outcomes

def toss_coin():
    r = random.randrange(2)
    if r == 0:
        return TAIL
    else:
        return HEAD

N = 20 # number of experiments to perform
for i in range(N):
    print ("experiment", i, "... outcome:", toss_coin())
```

Here's my output when I run the program:

```
[student@localhost book] python discrete-probability/tossfaircoin1.py
experiment 0 ... outcome: TAIL
experiment 1 ... outcome: HEAD
experiment 2 ... outcome: HEAD
experiment 3 ... outcome: TAIL
experiment 4 ... outcome: HEAD
experiment 5 ... outcome: HEAD
experiment 6 ... outcome: HEAD
experiment 7 ... outcome: HEAD
experiment 8 ... outcome: TAIL
experiment 9 ... outcome: HEAD
experiment 10 ... outcome: HEAD
experiment 11 ... outcome: HEAD
experiment 12 ... outcome: HEAD
experiment 13 ... outcome: HEAD
experiment 14 ... outcome: TAIL
experiment 15 ... outcome: TAIL
experiment 16 ... outcome: HEAD
experiment 17 ... outcome: TAIL
experiment 18 ... outcome: TAIL
experiment 19 ... outcome: HEAD
```

Much better isn't it? Not because I don't have a quarter, but because I can easily do a million coin-toss experiments in a split second by changing `N`. I can also collect all the outcomes and tabulate:

```
import random; random.seed()

HEAD = "HEAD"
TAIL = "TAIL"
S = [HEAD, TAIL] # outcomes

def toss_coin():
    if random.randrange(2) == 0:
        return TAIL
    else:
        return HEAD

N = 20 # number of experiments to perform
outcomes = []
for i in range(N):
    outcome = toss_coin()
    print ("experiment", i, "... outcome:", outcome)
    outcomes.append(outcome)
```

```
table = {}
table[HEAD] = outcomes.count(HEAD)
table[TAIL] = outcomes.count(TAIL)

print ("number of experiments:", N)
print ("number of heads:", table[HEAD])
print ("number of tails:", table[TAIL])
print ("probability of getting head:", table[HEAD] / float(N))
print ("probability of getting tail:", table[TAIL] / float(N))
```

Here's my output when I run the program:

```
[student@localhost book] python discrete-probability/tossfaircoin2.py
experiment 0 ... outcome: HEAD
experiment 1 ... outcome: HEAD
experiment 2 ... outcome: HEAD
experiment 3 ... outcome: TAIL
experiment 4 ... outcome: HEAD
experiment 5 ... outcome: TAIL
experiment 6 ... outcome: HEAD
experiment 7 ... outcome: TAIL
experiment 8 ... outcome: HEAD
experiment 9 ... outcome: TAIL
experiment 10 ... outcome: HEAD
experiment 11 ... outcome: HEAD
experiment 12 ... outcome: HEAD
experiment 13 ... outcome: HEAD
experiment 14 ... outcome: HEAD
experiment 15 ... outcome: TAIL
experiment 16 ... outcome: TAIL
experiment 17 ... outcome: HEAD
experiment 18 ... outcome: HEAD
experiment 19 ... outcome: HEAD
number of experiments: 20
number of heads: 14
number of tails: 6
probability of getting head: 0.7
probability of getting tail: 0.3
```

Exercise 9.1.1. Run the above program with a very large `N` (say 1000000). What is the probability of getting a head? Tail? No surprises, right? □

Of course if your `N` is really, really, really huge, you will find that the probability of getting a head is 0.5 and the probability of getting a tail is 0.5.

Here's a function (derived by simulation) that gives us the probability for each possible outcome of our experiment:

```
import random; random.seed()

HEAD = "HEAD"
TAIL = "TAIL"
S = [HEAD, TAIL] # outcomes

def toss_coin():
    if random.randrange(2) == 0:
        return TAIL
    else:
        return HEAD

N = 1000 # number of experiments to perform
outcomes = []
for i in range(N):
    outcome = toss_coin()
    #print "experiment", i, "... outcome:", outcome
    outcomes.append(outcome)

table = {}
table[HEAD] = outcomes.count(HEAD)
table[TAIL] = outcomes.count(TAIL)

print ("number of experiments:", N)
print ("number of heads:", table[HEAD])
print ("number of tails:", table[TAIL])

def p(x):
    return table[x] / float(N)

print ("probability of getting head:", p(HEAD))
print ("probability of getting tail:", p(TAIL))
```

I've increase `N` to 1000 and also commented out the printing of each experiment. Here's my output when I run the program:

```
[student@localhost book] python discrete-probability/tossfaircoin3.py
number of experiments: 1000
number of heads: 533
number of tails: 467
probability of getting head: 0.533
probability of getting tail: 0.467
```

Of course if we assume from the beginning that our coin is fair we can save the trouble of the computation:

```
import random; random.seed()

HEAD = "HEAD"
TAIL = "TAIL"
S = [HEAD, TAIL] # outcomes

def toss_coin():
    if random.randrange(2) == 0:
        return TAIL
    else:
        return HEAD

N = 1000 # number of experiments to perform
outcomes = []
for i in range(N):
    outcome = toss_coin()
    #print "experiment", i, "... outcome:", outcome
    outcomes.append(outcome)

table = {}
table[HEAD] = outcomes.count(HEAD)
table[TAIL] = outcomes.count(TAIL)

print ("number of experiments:", N)
print ("number of heads:", table[HEAD])
print ("number of tails:", table[TAIL])

def p(x):
    if x == HEAD: return 0.5
    elif x == TAIL: return 0.5
```

```
print ("probability of getting head:", p(HEAD))
print ("probability of getting tail:", p(TAIL))
```

Here's my output when I run the program:

```
[student@localhost book] python discrete-probability/tossfaircoin4.py
number of experiments: 1000
number of heads: 483
number of tails: 517
probability of getting head: 0.5
probability of getting tail: 0.5
```

Of course when `N` gets larger and larger, the probability function derived using simulations will match this new "theoretically" derived function.

Exercise 9.1.2. What if I change the `toss_coin` function to this:

```
def toss_coin():
    r = random.randrange(5)
    if r <= 1:
        return TAIL
    else:
        return HEAD
```

What is the probability of getting a head by simulations? Tail? What are the theoretical values? □

Exercise 9.1.3. Make a copy of the above program and then modify it to roll a fair die. Make 1000 rolls and report the number for each outcome. What is the simulated probability function? What is the theoretical probability function? □

Exercise 9.1.4. Write a program to simulate throwing two dice and returning the sum of their values. Of course the outcomes are 2, 3, 4, ..., 12. Are the probabilities for these outcomes the same? This is pretty helpful if you're gambling right? □

I abstract the coin toss experiment as follows: Let $S = \{\text{Head}, \text{Tail}\}$ represent all possible outcomes of my coin toss experiment. S is the called the **sample space** of my random experiment. Head and Tail are called the

sample space

outcomes of my experiment. I define a function

outcomes

$$p : S \to \mathbb{R}$$

such that $p(\text{Head})$ gives us the chance that a coin toss will give us a head and $p(\text{Tail})$ gives us the chance of getting a tail. So if the coin is fair, I would have

$$p(\text{Head}) = 1/2 = p(\text{Tail})$$

p is called a **probability distribution function**.

probability distribution function

9.1.2 Die rolls

Look at a die. When you roll a die, you get the face value of the roll, i.e., the number of dots on the top surface of the die, you get one of the following:

$$\texttt{ONE}, \texttt{TWO}, \texttt{THREE}, \texttt{FOUR}, \texttt{FIVE}, \texttt{SIX},$$

Assuming that the die is fair and you roll the die a huge number of times, you will see that approximately 1/6 of the rolls will give you a `ONE` (and likewise for the other cases.)

I can assign 1/6 to `ONE` to indicate that the proportion of rolls that gives me `ONE`. Likewise for the other cases. This is basically a function:

$$\begin{aligned} p(\texttt{ONE}) &= 1/6 \\ p(\texttt{TWO}) &= 1/6 \\ p(\texttt{THREE}) &= 1/6 \\ p(\texttt{FOUR}) &= 1/6 \\ p(\texttt{FIVE}) &= 1/6 \\ p(\texttt{SIX}) &= 1/6 \end{aligned}$$

Such a function is of the form

$$p : \{\texttt{ONE}, \texttt{TWO}, \texttt{THREE}, \texttt{FOUR}, \texttt{FIVE}, \texttt{SIX}\} \to [0, 1]$$

Another thing to note is that

$$p(\texttt{ONE}) + p(\texttt{TWO}) + p(\texttt{THREE}) + p(\texttt{FOUR}) + p(\texttt{FIVE}) + p(\texttt{SIX}) = 1$$

In general, the set $\{\texttt{ONE}, \texttt{TWO}, \texttt{THREE}, \texttt{FOUR}, \texttt{FIVE}, \texttt{SIX}\}$ is called the set of **outcomes**, or the **sample space**, of my experiment of rolling a die.

outcomes
sample space

For us, our sample space is a finite set (you can also talk about probability theory for non-finite sets.)

More generally, given a finite set of outcomes of performing an experiment, say S, a **probability function** on S is just a function of the form

probability function

$$p : S \to [0, 1]$$

such that

$$\sum_{x \in S} p(x) = 1$$

A probability function is also called a **probability distribution function** because it distributes the maximum probability of 1 to various outcomes.

probability distribution function

Exercise 9.1.5. Dr. Liow has a biased die: `SIX` appears half the time (when you repeat the experiment of rolling it for a large number of times.) What is $p(x)$ for $x \neq$ `SIX`.

Exercise 9.1.6. Dr. Liow has a bent dime: `HEAD` appears three times more than `TAIL`. What is the sample space for the experiment of tossing the coin? (Duh). What is $p($`HEAD`$)$? What is $p($`TAIL`$)$?

Exercise 9.1.7. A device generates $\{0,1\}^4$, i.e., bit strings of length 4 uniformly. When I say "uniformly", I mean the outcomes are all equally likely. What are all the outcomes? What is the probability of each outcome?

A probability distribution function is **uniform** if $p(x)$ is the same for all x in the sample space of the function. When I say that I'm rolling a **fair** die, I mean that the probability function for the experiment (of rolling the die) is uniform.

uniform

fair

Instead of asking the probability of getting a `SIX` when I roll a die, I might be interested in getting `ONE` or `SIX`. In general, a subset A of the set of all outcomes is called an **event**. In that case, I will write

$$p(A)$$

or

$$p[A]$$

for

$$\sum_{x \in A} p(x)$$

So for instance in the case of our die,

$$p[\{\texttt{ONE}, \texttt{SIX}\}] = p(\texttt{ONE}) + p(\texttt{SIX}) = 1/6 + 1/6 = 1/3$$

Let A and B be events of an experiment with sample space S and probability distribution function p. Then

$$p[A \cup B] \leq p[A] + p[B]$$

This means that if we know a lot about the probability of A and probability of B, but we're not very sure about the common events of A and B, we can still bound the probability of $A \cup B$. If we do know something about the probability of $A \cap B$, then we have

$$p[A \cup B] = p[A] + p[B] - p[A \cap B]$$

If $A \cap B = \emptyset$, we get

$$p[A \cup B] = p[A] + p[B]$$

File: discrete-probability/experimental-approach.tex

9.2 Experimental (or Empirical) Approach

Here's how to think of probability intuitively: You can of course think of probability as some kind of averaging. In the case of tossing a particular coin, if you say

$$p(\text{Head}) = \frac{1}{3}$$

what you meant is that if you toss the coin 3 times, then it's likely that approximately 1 out of the 3 is a head; if you toss the coin 300 times, then it's very likely that approximately 100 of the 300 are heads; if you toss the coin 3000 times, then it's very very likely that approximately 1000 of the 3000 are heads; etc. The more tosses you make, the closer you get to "one third of the tosses comes out head".

In the above coin toss experiment, I have two possible outcomes: either I get a head or I get a tail. I create two symbols to represent these two possible outcomes: Head and Tail. The names is arbitrary and it's entirely up to you to come up with symbols for all the possible outcomes. For instance you can name the outcomes H and T instead of Head and Tail.

How let's formalize the notation for studying probability theory ...

File: discrete-probability/probability-distribution-function.tex

9.3 Probability distribution function

A (finite) discrete **probability distribution function** is a real-valued function on a finite set S

probability distribution function

$$p : S \to \mathbb{R}$$

such that

[DP-1] $0 \leq p(x) \leq 1$ for all $x \in S$
[DP-2] $\sum_{x \in S} p(x) = 1$

Because of [DP-1], I can also write p like this:

$$p : S \to [0, 1]$$

For simplicity, if A is a subset of S, I'll write

$$p(A) = \sum_{x \in A} p(x)$$

[DP-2] is the same as saying

$$p(S) = 1$$

In the above context, S is a **sample space**. the elements of S are called **outcomes**, and a subset of S is called an **event**.

sample space
outcomes
event

The **Euler totient function** ϕ is defined to as follows: For $n > 0$, $\phi(n)$ is the number of integers k such that $1 \leq k \leq n$ and $\gcd(k, n) = 1$.

Euler totient function
ϕ

Here are some basic facts about probability functions.

Theorem 9.3.1. *Let $p : S \to [0, 1]$ be a probability distribution function.*

(a) *If $A \subseteq S$, then $0 \leq p(A) \leq 1$.*

(b) $p(\emptyset) = 0$.

(c) *If A and B are subsets of S, then*

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

(d) *If A and B are subsets of S, then*

$$p(A \cup B) \leq p(A) + p(B)$$

(e) *If A and B are disjoint subsets of S, then*

$$p(A \cup B) = p(A) + p(B)$$

(f) *If $A \subseteq S$, then $p(\overline{A}) = 1 - p(A)$.*

Exercise 9.3.1. Prove all the statements in the above theorem. □

Definition 9.3.1. A probability distribution function $p : S \to [0, 1]$ is said to be **uniform** if

uniform

$$p(x) = \frac{1}{|S|}$$

for each $x \in S$. $|S|$, the size of S, denotes the numbers of distinct elements in S.

Example 9.3.1. Suppose that I've rigged my coin so that the chance of getting a head is twice of getting the tail. This means that the probability function

$$p : S \to [0, 1]$$

is

$$p(\text{Head}) = \frac{2}{3}, \quad p(\text{Tail}) = \frac{1}{3}$$

Of course if the coin is a fair coin I would have

$$p(\text{Head}) = \frac{1}{2} = p(\text{Tail})$$

□

Exercise 9.3.2. Consider the random experiment of rolling a fair die, except that I'm only interested in two outcomes of "I get a six" and "I did not get a six". Write down this pdf. □

Example 9.3.2. Suppose I have a loaded die such that getting a "1" is 10 times more likely than the others and the others are equally likely. Let p be

the probability function of rolling this die. Then

$$\begin{aligned} 1 &= p(\text{One}) + p(\text{Two}) + \cdots + p(\text{Six}) \\ &= p(\text{One}) + 5 \cdot p(\text{Two}) \\ &= p(\text{One}) + 5 \cdot \frac{1}{10} p(\text{One}) \\ &= \frac{3}{2} \cdot p(\text{One}) \end{aligned}$$

So

$$p(\text{One}) = \frac{2}{3}$$

and

$$p(\text{Two}) = p(\text{Three}) = ... = p(\text{Six}) = \frac{2}{30}$$

And of course

$$S = \{\text{One}, \text{Two}, \text{Three}, \text{Four}, \text{Five}, \text{Six}\}$$

Go ahead and write a program to simulate tossing my loaded die. Run it for a large number of experiments and check that you get the expected results.

```
import random; random.seed()

ONE = "ONE"
TWO = "TWO"
THREE = "THREE"
FOUR = "FOUR"
FIVE = "FIVE"
SIX = "SIX"

# write a probability function for this experiment

def roll_die():
    r = random.randrange(30)
    if r < 20: return ONE
    elif r < 22: return TWO
    elif r < 24: return THREE
    elif r < 26: return FOUR
    elif r < 28: return FIVE
    else: return SIX

NUM_EXPERIMENTS = 20
for i in range(NUM_EXPERIMENTS):
    print ("experiment", i, "... outcome:", roll_die())
```

The following function can be used to produce the different experiment functions (you should type it up and run it ... if you're too lazy to do that you can also grab the code from our class web site ... look for "probability library"):

```
def build_experiment(xs):
    total = sum([b for (a,b) in xs])
    accumulate = xs[0]
    for (a,b) in xs[1:]:
        accumulate.append((a, accumulate[-1][1] + b)
    def experiment():
        r = random.randrange(total)
        for (a,b) in accumulate:
            if r < b: return a
        raise ValueError("r:%s greater than total:%s" %\
                         (r, total))
    return experiment

HEAD = "HEAD"
TAIL = "TAIL"
toss_coin = build_experiment([(HEAD, 2),
                              (TAIL, 1),
                             ])
ONE = "ONE"
TWO = "TWO"
THREE = "THREE"
FOUR = "FOUR"
FIVE = "FIVE"
SIX = "SIX"
roll_die = build_experiment([(ONE, 20),
                             (TWO, 2),
                             (THREE, 2),
                             (FOUR, 2),
                             (FIVE, 2),
                             (SIX, 2),
                            ])
```

Exercise 9.3.3. The tetris game has 5 shapes: SQUARE, l, Z, S, T. To make your tetrix game annoying, you have decided to make Z 5 times more likely to appear than S, S 5 times more likely than T, T 5 times more like than l, and l 5 times more likely than SQUARE. (And you make the shapes fall down real fast too.) Write down the probabilities of the shapes. You gave your game to Dr. Liow who (promptly) dies after playing the game for 20 seconds. If a total of 100 shapes appear, approximately how many of each shape appear.

Exercise 9.3.4. The probability of getting a quiz from Dr. Liow during a regular class meeting is 0.8. What is the probability of getting 3 quizzes in a row? Note that what is given is the experiment "Attend a regular class meeting". (analogous to "toss a coin") The sample space is

$$S = \{\text{Quiz}, \text{NoQuiz}\}$$

(analogous to $S = \{\text{Head}, \text{Tail}\}$.) Let's shorten that to

$$S = \{\text{Q}, \text{N}\}$$

Now what is asked is actually related to another experiment: "Attend 3 classes in a row". (Later you'll see how to view this as a "product of experiments".) There are two ways to compute an approximation of the desired probability.

FIRST METHOD: You should use the first experiment to generate a sequence of Q's and N's:

$$\text{Q}, \text{Q}, \text{N}, \text{Q}, \text{Q}, \text{Q}, \text{Q}, \text{Q}, \text{Q}, \text{N}, \text{Q}, \text{Q}, ...$$

satisfying the given probability that $p(\text{Q}) = 0.8$ and $p(\text{N}) = 0.2$. Next you look at all the consecutive triples. For instance the above gives

$$\text{QQN}, \text{QNQ}, \text{NQQ}, \text{QQQ}, \text{QQQ}, \text{QQQ}, \text{QQQ}, \text{QQQ}, \text{QQN}, \text{QNQ}, \text{NQQ}, ...$$

and then compute the approximate probability of seeing QQQ in this sampling. Of course you need to have a reasonbly huge sample. You can either do this by hand or write a program to simulate the experiment.

SECOND METHOD: If you have a function for the first experiment, say `attend_class` which gives you Q or N, then you should have a function say `attend_3_classes` that uses `attend_class`:

```
Q = 'Q'
N = 'N'

def attend_3_classes():
    return [attend_class(),
            attend_class(),
            attend_class(),
           ]
```

Of course you can now generate samples and count:

```
MAX = 100000
count = 0
for i in range(MAX):
    if attend_3_classes() == [Q, Q, Q]:
        count += 1
print (float(count) / MAX)
```

Exercise 9.3.5. John Cantcode is a pretty bad programmer. At Fullabugz, it is known that the chance of him writing a buggy C++ program is 3 times that of his Python program. (Fullabugz uses only two languages. Why? More than two means more headaches for John's manager.) One third of the programs at Fullabugz is C++ programs and according to statistics the percentage of C++ programs that crashed in the past 10 years at the company is 80%. John writes 1/4 of all the programs at Fullabuz. His manager just ran a program and it crashed. What is the probability that this was a C++ program? What is the probability that this program was written by John? (Do this empirically, i.e., do some simulations.) Is there enough information to compute the probability? If not, what is lacking?

File: discrete-probability/rv.tex

9.4 Random variable

Now for the definition of a very confusing term: random variable. A **random variable** X of S (a sample space) is just a function from S to some set, say V. random variable

$$X : S \to V$$

Usually V is the set of real numbers:

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

It's really important to remember that a random variable is not random and is not a variable!!! It's just a function from a sample space to a set. This is an example of a badly chosen name for this idea.

There are two main reasons why we need the concept of random variables. Pay attention to the following.

9.4.1 Random variable as labels

The first reason for having random variables is that we want to create some labels for the values in a sample space. The labels will then create subsets of the sample space, i.e., random variables helps create meaningful events.

As an example, I'm going back to the die roll experiment. The sample space is

$$S = \{\text{One}, \text{Two}, \text{Three}, \text{Four}, \text{Five}, \text{Six}\}$$

Suppose I'm playing a game with a die and I win if the roll gives me either One or Six; otherwise I lose. I can then define the following random variables:

$$X : \{\text{One}, ..., \text{Six}\} \to \{\text{Good}, \text{Bad}\}$$

where

$$X(\text{One}) = X(\text{Six}) = \text{Good}$$
$$X(\text{Two}) = X(\text{Three}) = X(\text{Four}) = X(\text{Five}) = \text{Bad}$$

We have create two subsets

$$A = \{s \in S \mid X(s) = \text{Good}\}$$
$$B = \{s \in S \mid X(s) = \text{Bad}\}$$

of S. In other words, random variables create events for each value in the range of the random variables. In fact the subsets created are disjoint and the union of all these subsets for the original sample space. In other words, X creates a partition of the sample space, i.e., A and B are disjoint and $A \cup B$ is X.

As a consequence, my random variable X also defines a new probability distribution function

$$p_X : \{\text{Good}, \text{Bad}\} \to [0, 1]$$

where

$$p_X(\text{Good}) = p(\{s \in S \mid X(s) = \text{Good}\}) = 1/3$$
$$p_X(\text{Bad}) = p(\{s \in S \mid X(s) = \text{Bad}\}) = 2/3$$

This function is very frequently written like this:

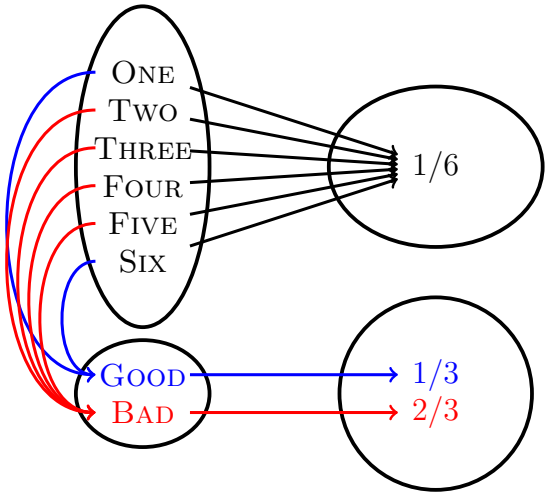
$$\Pr[X = \text{Good}] = 1/3$$
$$\Pr[X = \text{Bad}] = 2/3$$

or sometimes $\text{P}[X = \text{Good}]$ or $\text{P}(X = \text{Good})$. This Pr is a bad notation because this function actually depends on p and X. But nobody write $\Pr_p[X =$

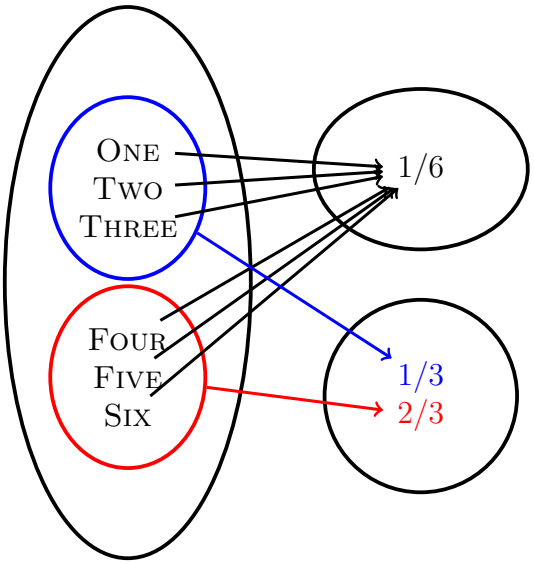
Good]. Part of the reason is because a random variable X is usually tied to a probability distribution function. Which is not exactly true since X is a function of the sample space and is therefore tied to the sample space and not the probability function itself. But this is the practice and you have to remember what I said above in order not to be confused.

It's important to note that p_X and $\Pr[X = \bullet]$ are pdfs, but they are *not* pdf on S: they are pdfs on $\{\text{Good}, \text{Bad}\}$.

Here's a picture to keep in mind:



The function at the top is p and the function at the bottom is $\Pr[X = \bullet]$. You can (and should) think of a random variable as collecting probabilities given by p into buckets. You can think of it this way:



The random variable X basically collects up probabilities: One and Six are collected up into Goodand the rest are collected up into Bad.

It's possible, in fact it's common, to have multiple random variables on the same sample space.

Here's another random variable:

$$Y : S \to \{\text{Even}, \text{Odd}\}$$

where

$$Y(\text{One}) = Y(\text{Three}) = Y(\text{Five}) = \text{Odd}$$
$$Y(\text{Two}) = Y(\text{Four}) = Y(\text{Six}) = \text{Even}$$

This means that we can talk about

$$\Pr[Y = \text{Odd}]$$

which of course is just

$$p(\{s \in S \mid Y(s) = \text{Odd}\})$$

Our Y allows us to create the following subsets of S:

- $\{x \in S \mid Y(x) = \text{Even}\} = \{\text{Two}, \text{Four}, \text{Six}\} \subseteq S$
- $\{x \in S \mid Y(x) = \text{Odd}\} = \{\text{One}, \text{Three}, \text{Five}\} \subseteq S$

This Y creates a pdf

$$\Pr[Y = \bullet] : \{\text{Even}, \text{Odd}\} \to [0, 1]$$

where

$$\Pr[Y = \text{Even}] = \Pr[Y = \text{Odd}] = \frac{1}{2}$$

Here's yet another random variable on die rolls:

$$Z : S \to \{\text{Small}, \text{Medium}, \text{Large}\}$$

where

$$\begin{aligned} Z(\text{One}) = Z(\text{Two}) &= \text{Small} \\ Z(\text{Three}) = Z(\text{Four}) &= \text{Medium} \\ Z(\text{Five}) = Z(\text{Six}) &= \text{Large} \end{aligned}$$

Frequently instead of just

$$\Pr[Z = \text{Small}]$$

you actually see expressions like this:

$$\Pr[Z = \text{Small or } Z = \text{Medium}]$$

or more formally

$$\Pr[Z \in \{\text{Small}, \text{Medium}\}]$$

In other words

$$\Pr[\text{boolean expression on a random variable}]$$

and even boolean expressions involving more than one random variables:

$$\Pr[Y = \text{Odd} \vee Z = \text{Small}]$$

Here's another random variable where the value of the random variable takes a real values. For instance, for the die rolling experiment, suppose I define W in the obvious way:

$$\begin{aligned} W(\text{One}) &= 1 \\ W(\text{Two}) &= 2 \\ W(\text{Three}) &= 3 \\ W(\text{Four}) &= 4 \\ W(\text{Five}) &= 5 \\ W(\text{Six}) &= 6 \end{aligned}$$

With this random variable W, I can write

$$\Pr[W \text{ is odd}]$$

to mean $\Pr[W \in \{\text{One}, \text{Three}, \text{Five}\}]$. I can even write $\Pr[W \leq 4]$. The meaning should be obvious.

In general, suppose $p : S \to \mathbb{R}$ is a probability distribution function and $X : S \to V$ is a random variable. If $v \in V$. I will write $\Pr[X = v]$ for the following:

$$\Pr[X = v] = p(\{s \in S \mid X(s) = v\}) = p(X^{-1}(v))$$

This function

$$\Pr[X = \bullet] : V \to [0, 1]$$

is a pdf. In other words, you can get a pdf from a pdf and a random variable X.

WARNING: In some books, instead of

- *Let p be the probability function of tossing a coin that is twice as likely to get a head than a tail, then*

$$p : \{\text{Head}, \text{Tail}\} \to [0, 1]$$

$$p(\text{Head}) = \frac{2}{3}, \quad p(\text{Tail}) = \frac{1}{3},"$$

you might hear this:

- *Let X be the random variable of tossing a coin that is twice as likely to get a head than a tail, then*

$$\Pr[X = \text{Head}] = \frac{2}{3}, \quad \Pr[X = \text{Tail}] = \frac{1}{3}$$

Treat them as the same. Some authors do not differentiate between probability distribution function and the probability distribution derived from a random variable. This practice is very common, so you have to learn to live with it.

9.4.2 Random variable as a scoring function; expectation

Let $p : S \to [0, 1]$ be a pdf. Besides labels, you can also think of the random variable X on S as some kind of *scoring* for each outcome of the sample space of an experiment:

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

With this setup, we can define the average value or the **expected value** of X, or the **expectation** of X, to be

expected value

expectation

$$\mathrm{E}[X] = \sum_{s \in S} X(s) \cdot p(s)$$

In this case, $\mathrm{E}[X]$ is the average score you will get if you keep drawing some outcome from S. What I mean is this:

Suppose you close your eyes and randomly draw an s_0 from S and note down the score of s_0, i.e., $X(s_0)$. You put then s_0 back into S. You repeat the above and get $s_1 \in S$ and note down the score $X(s_1)$. You put s_1 back into S. The average so far is $(X(s_0) + X(s_1))/2$. You repeat the above and get $s_2 \in S$. The average is now $(X(s_0) + X(s_1) + X(s_2))/3$. You keep on going and you'll get the theoretical average. ("Theoretical" because obviously you can't go on forever.) The value of $\mathrm{E}[X]$ gives you this theoretical average.

Instead of summing over X, it's also possible to sum over all the possible values of X:

$$\mathrm{E}[X] = \sum_{x \in X(S)} x \cdot \Pr[X = x]$$

See example below if you don't see why. Here, $X(S)$ is the range of X, i.e.,

$$X(S) = \{X(s) \mid s \in S\}$$

In many cases, the probability of an experiment is not what you're after. It's usually some kind of "gain" or "value" from doing an experiment or playing a probabilistic game – which turns out to be the expectation of some random variable. This is what I mean:

Take a look at our random variable X:

$$X : \{\textsc{One}, ..., \textsc{Six}\} \to \{\textsc{Good}, \textsc{Bad}\}$$

where

$$X(\textsc{One}) = X(\textsc{Six}) = \textsc{Good}$$
$$X(\textsc{Two}) = X(\textsc{Three}) = X(\textsc{Four}) = X(\textsc{Five}) = \textsc{Bad}$$

Suppose I now define another random variable:

$$X' : \{\text{One}, ..., \text{Six}\} \to \mathbb{R}$$

$$\begin{aligned} X'(\text{One}) = X'(\text{Six}) &= 3 \\ X'(\text{Two}) = X'(\text{Three}) = X'(\text{Four}) = X'(\text{Five}) &= -1 \end{aligned}$$

This could be for instance a gambing game where I get \$3 if I roll a one or a six; otherwise I lose \$1. The expectation of X' is

$$\mathrm{E}[X'] = \sum_{s \in S} X'(s)p(x)$$

which is

$$\begin{aligned} \mathrm{E}[X'] &= X'(\text{One})p(\text{One}) + X'(\text{Two})p(\text{Two}) + X'(\text{Three})p(\text{Three}) \\ &\quad + X'(\text{Four})p(\text{Four}) + X'(\text{Five})p(\text{Five}) + X'(\text{Six})p(\text{Six}) \\ &= 3 \cdot \frac{1}{6} + (-1) \cdot \frac{1}{6} + (-1) \cdot \frac{1}{6} + (-1) \cdot \frac{1}{6} + (-1) \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} \\ &= \frac{1}{3} \end{aligned}$$

which tells me on the average I'm going to gain \$1/3 dollars, about 33 cents, per game.

Note that I can also calculate the probability distribution from X':

$$\Pr : \{3, -1\} \to [0, 1]$$

to get

$$\begin{aligned} \Pr[X' = 3] &= p(\{\text{One}, \text{Six}\}) = 1/3 \\ \Pr[X' = -1] &= p(\{\text{Two}, \text{Three}, \text{Four}, \text{Five}\}) = 2/3 \end{aligned}$$

and then compute $\mathrm{E}[X']$ this way:

$$\begin{aligned} \mathrm{E}[X'] &= \sum_{x \in X'(S)} x \cdot \Pr[X' = x] \\ &= 3 \cdot \Pr[X' = 3] + (-1) \cdot \Pr[X' = -1] \\ &= 3 \cdot \frac{1}{3} + (-1)\frac{2}{3} \\ &= \frac{1}{3} \end{aligned}$$

In general, if X is a random variable with values in $\mathbb{R}$, then

$$\mathrm{E}[X] = \sum_{s \in S} X(s) \cdot p(x) = \sum_{x \in X(S)} x \cdot \Pr[X = x]$$

Suppose for this gambling game, to play the game, you have to pay \$0.5 (fifty cents) for each roll. You can work out $\mathrm{E}[X'] = 1/3$ and then subtract 0.5 to get your gain per game to get a gain of

$$\frac{1}{3} - 0.5 = -\frac{1}{6}$$

Another way to do that is to include the cost of playing the game in the expectation computation. For instance you can define a random variable

$$Y' : \{\textsc{One}, \textsc{Two}, \textsc{Three}, \textsc{Four}, \textsc{Five}, \textsc{Six}\} \to \mathbb{R}$$

as

$$Y'(s) = -0.5$$

Then the gain from playing this game (including the cost of playing the game is

$$\mathrm{E}[X' + Y']$$

where $X' + Y'$ is the random variable

$$X' + Y' : \{\textsc{One}, \textsc{Two}, \textsc{Three}, \textsc{Four}, \textsc{Five}, \textsc{Six}\} \to \mathbb{R}$$

given by

$$(X' + Y')(s) = X'(s) + Y'(s)$$

If you compute $\mathrm{E}[X' + Y']$ using $\mathrm{E}[X' + Y'] = \sum_{s \in S}(X(s) + Y(s))p(s)$, you will get

$$(3{-}0.5)\frac{1}{6}{+}(-1{-}0.5)\frac{1}{6}{+}(-1{-}0.5)\frac{1}{6}{+}(-1{-}0.5)\frac{1}{6}{+}(-1{-}0.5)\frac{1}{6}{+}(3{-}0.5)\frac{1}{6} = -\frac{1}{6}$$

If I use $\mathrm{E}[X' + Y'] = \sum_{z \in (X+Y)(S)} z \cdot \Pr[X + Y = z]$. For this I'll need to know the range of $X + Y$:

$$(X + Y)(S) = \{3 - 0.5, -1 - 0.5\} = \{2.5, -1.5\} = \{2.5, -1.5\}$$

and the probability distribution function of $\Pr[X + Y = \bullet]$ is

$$\begin{aligned}
\Pr[X + Y = 2.5] &= \{s \in S \mid (X + Y)(s) = 2.5\} \\
&= p(\{\textsc{One}, \textsc{Six}\}) \\
&= \frac{2}{6} = \frac{1}{3} \\
\Pr[X + Y = -1.5] &= \{s \in S \mid (X + Y)(s) = -1.5\} \\
&= p(\{\textsc{Two}, \textsc{Three}, \textsc{Four}, \textsc{Five}\}) \\
&= \frac{4}{6} = \frac{2}{3}
\end{aligned}$$

Therefore

$$\begin{aligned}\mathrm{E}[X'+Y'] &= \sum_{z\in(X'+Y')(S)} z\cdot\Pr[X'+Y'=z]\\ &= 2.5\cdot\frac{1}{3}+(-1.5)\cdot\frac{2}{3}\\ &= -\frac{1}{6}\end{aligned}$$

You can also use the formula

$$\mathrm{E}[X'+Y'] = \mathrm{E}[X'] + \mathrm{E}[Y']$$

I already know that $\mathrm{E}[X'] = 1/3$. $\mathrm{E}[Y']$ is easy:

$$\mathrm{E}[Y'] = (-0.5)\frac{1}{6}+\cdots+(-0.5)\frac{1}{6} = (-0.5)(1) = -0.5$$

Therefore $\mathrm{E}[X'+Y'] = 1/3 - 0.5 = -1/6$.

This is important: Look at the three ways of computing $\mathrm{E}[X'+Y']$ again. Which is easier? And, more importantly, why is the one that you picked the easiest?

You'll find that computing $\mathrm{E}[X'+Y']$ using

$$\mathrm{E}[X'+Y'] = \mathrm{E}[X'] + \mathrm{E}[Y']$$

is frequently the easiest.

In general if X and Y are random values:

$$\begin{aligned}X &: S\to\mathbb{R}\\ Y &: S\to\mathbb{R}\end{aligned}$$

then of course you can also consider

$$X+Y : S\to\mathbb{R}$$

This is clearly also a random variable. Remember that random variables are just functions. More generally, if there are n random variables

$$X_i : S\to\mathbb{R}$$

for $i = 0, 1, 2, \ldots, n-1$, then

$$\sum_{i=0}^{n-1} X_i : S\to\mathbb{R}$$

where

$$\left(\sum_{i=0}^{n-1} X_i\right)(s) = \sum_{i=0}^{n-1} X_i(s)$$

is also a random variable. Furthermore

$$\mathrm{E}\left[\sum_{i=0}^{n-1} X_i\right] = \sum_{i=0}^{n-1} \mathrm{E}\left[X_i\right]$$

At this point, it should not be surprising that if I change the rules of the above gambling game so that the random variable is $X'' = 2X'$, i.e., if you roll a one or a six you win $2 \cdot 3$ dollar otherwise you lose $2 \cdot 1$ dollars, then the average gain per game should be

$$\mathrm{E}[X''] = \mathrm{E}[2X'] = 2 \cdot \mathrm{E}[X']$$

Right?

In general, if a is a real number and $X : S \to \mathbb{R}$ is a random variable, then $aX :\to \mathbb{R}$ given by

$$(aX)(s) = a \cdot X(s)$$

is also a random variable and furthermore

$$\mathrm{E}[aX] = a \cdot \mathrm{E}[X]$$

Exercise 9.4.1. If $X : S \to \mathbb{R}$ is a random variable and $a \in R$, prove that

$$\mathrm{E}[aX] = a \cdot \mathrm{E}[X]$$

□

It's also common to view a real number, say b, as a random variable. What I mean is this random variable:

$$b : S \to \mathbb{R}$$

where

$$b(s) = b$$

for any outcome s in S. In other words, I'm consider b as a constant function. For instance our

$$Y' : \{\textsc{One}, \textsc{Two}, \textsc{Three}, \textsc{Four}, \textsc{Five}, \textsc{Six}\}$$

above where

$$Y'(s) = -0.5$$

(i.e., you have to pay 50 cents for every roll) is such as example. So instead of defining Y' and then say

$$\mathrm{E}[X' + Y'] = -\frac{1}{6}$$

I could have just said

$$\mathrm{E}[X' - 0.5] = \mathrm{E}[X'] + \mathrm{E}[-0.5] = \frac{1}{3} + (-0.5) = -\frac{1}{6}$$

(i.e., Y' is just the constant function $Y' = -0.5$.) Easy right?

Let me collection the above together as a theorem:

Theorem 9.4.1. *For $i = 0, 1, 2, ..., n-1$, let X_i be a (real-valued) random variable and c_i be a constant. Let c be a constant (viewed as a constant random variable). Then*

$$\mathrm{E}\left[\sum_{i=0} c_i X_i\right] = \sum_{i=0} c_i \,\mathrm{E}\left[X_i\right]$$

and

$$\mathrm{E}[c] = c$$

Exercise 9.4.2. You should prove that if b is a real number and b also represent the random variable $S \to \mathbb{R}$, then

$$\mathrm{E}[b] = b$$

(The b in $\mathrm{E}[b]$ is a random variable. The b on the right of the above equation is a real number. Correct? Make sure read your math carefully.) □

Exercise 9.4.3. Is it true that if X and Y are random variables, then $\mathrm{E}[XY] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$? Try a couple of examples by hand and see what you can discover. □

9.4.3 Labeling and scoring

Frequently, a random variable is used for both labeling and scoring. For instance

$$X' : \{\text{One}, \text{Two}, \text{Three}, \text{Four}, \text{Five}, \text{Six}\} \to \mathbb{R}$$

with

$$X'(\text{One}) = X'(\text{Six}) = 3$$
$$X'(\text{Two}) = X'(\text{Three}) = X'(\text{Four}) = X'(\text{Five}) = -1$$

X' scores the outcomes, but X' (obviously) also label outcomes based on their scores.

Exercise 9.4.4. Consider the random experiment of rolling two fair dice. The sample space is $S = S_1 \times S_2$ where $S_i = \{\text{One}, \text{Two}, \ldots, \text{Six}\}$ $(i = 0, 1)$. For instance $(\text{One}, \text{Two}) \in S$ is the outcome that the first die is One and the second die is Two. The probability is uniform. Let $X_1((s_1, s_2)) = s_1$ and $X_2((s_1, s_2)) = s_2$.

- What is the probability of rolling two sixes? Before you write down the probability, I want you to write the "probability of rolling two sixes" using the Pr notation using the random variables X_1 and X_2.
- What is the probability of rolling a sum of 10? Before you write down the probability, I want you to write the "probability of rolling a sum of 10" using the Pr notation using a random variable that you have to define on your own.

□

Exercise 9.4.5. Let `x[0..9]` be an array containing a random permutation of $0, 1, 2, ..., 9$.

- What is the probability that `x[0]` is `0`? Write a C++ program to perform a random experiment to prove what you just said. Can you prove what you just said?
- Let `i` be an integer in `[0, 9]`. What is the probability that `x[i]` is `i`? Write a C++ program to perform a random experiment to verify what you just said. Can you prove what you just said.
- Let `i` and `j` be two distinct integers in `[0, 9]`. What is the probability

that `x[i]` is `i` and `x[j]` is `j`? Write a C++ program to perform a random experiment to verify what you just said. Now prove what you just said.

- What is average number of `i` such that `x[i]` is `i`? Write a C++ program to perform a random experiment to verify what you just said. Can you prove what you just said?

□

File: discrete-probability/indicator.tex

9.5 Indicator random variable

Suppose you have a sample space S and A is an event of S, i.e., $A \subseteq S$. The following is a very useful random variable:

$$X_A : S \to \mathbb{R}$$

where

$$X_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{otherwise} \end{cases}$$

This is like a labeling: values in A are labeled with 1 while values not in A are labeled with 0. Such a random variable is called an **indicator random variable**. It's also common to write I_A for such a random variable.

indicator random variable

There are times when A has only a single value. Say $a \in S$. Then I will write X_a instead of $X_{\{a\}}$. In other words, the indicator random variable for a is

$$X_a : S \to \mathbb{R}$$

where

$$X_a(s) = \begin{cases} 1 & \text{if } s = a \\ 0 & \text{otherwise} \end{cases}$$

You can think of X_A as a boolean function that tells you if an outcome falls in A. Another way is to think of X_A as a counter that counts (or labels) an outcome as 1 if the outcome is in A. This is a very important way to think about the indicator random variable especially when we do expected value computations. For instance, suppose you consider a random experiment of tossing a coin. The sample space is $\{\text{Head}, \text{Tail}\}$. The indicator variable X_{Head} counts the number of heads: it's either zero or one.

Exercise 9.5.1. Let X_A be an indicator random variable on $A \subseteq S$. Prove that

$$\mathrm{E}[X_A] = p(A)$$

(This is very useful and will be used frequently in examples and exercises.)

Proof. Exercise. □

File: discrete-probability/conditional-probability.tex

9.6 Conditional Probability

In probability theory, you might see something like "what is the chance of getting a 1 from this die?" Another type of probability question looks like this: "What is the chance of getting a 1 *if the result is odd*?" Or "Suppose I throw two dice. What is the chance that the second gives a 5 if the first gives a 1?"

In the real world, you might hear a question like: "What is the likelihood of getting cancer if I'm a smoker?" (which is not the same as the likelihood of getting cancer in general) or "What is the chance of being runned down by a truck if I'm shortsighted and there's a 50-50 chance that I forgot my glasses?" (which is not the same as the chance of being runned down by a truck in general) or "What is the probability that my laptop will catch a virus if I only update my anti-virus definition once a month?" (which is not the same as the chance of having a malfunctioning laptop for the general person)

Here's how you think of it.

Suppose I say that "The chance of getting a 1 when I roll this die *if the result is odd* is $\frac{1}{3}$." What I mean is this: If I roll the die n times where n is huge, then out of all the results which are odd (i.e., I ignore the even results), then about $\frac{1}{3}$ of them are 1.

In notation, I might write

$$p(A \mid B)$$

where $A = \{\text{One}\}$ and $B = \{\text{One}, \text{Three}, \text{Five}\}$, i.e.,

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

This assumes that $p(B)$ is not 0.

Note the difference between $p(A \cap B)$ and $p(A \mid B)$. To understand the difference intuitively, think about the difference between "What is the chance of rain and thunderstorm today?" and "What is the chance of rain if there is thunderstorm today?"

Exercise 9.6.1. For the case of rolling a die, $A = \{\text{One}\}$, and $B = \{\text{One}, \text{Three}, \text{Five}\}$, compute $p(A \mid B)$ and $p(A \cap B)$.

Consider another scenario: Suppose I ask "What is the chance of getting of getting a small number (i.e., either 1 or 2) if the result is odd?" In that case $A = \{\text{One}, \text{Two}\}$ and $B = \{\text{One}, \text{Three}, \text{Five}\}$. The quantity I need is

$$\begin{aligned} p(A \mid B) &= \frac{p(A \cap B)}{p(B)} \\ &= \frac{p(\{\text{One}, \text{Two}\} \cap \{\text{One}, \text{Three}, \text{Five}\})}{p(\text{One}, \text{Three}, \text{Five})} \\ &= \frac{p(\{\text{One}\})}{p(\{\text{One}, \text{Three}, \text{Five}\})} \\ &= \frac{1/6}{1/2} \\ &= \frac{1}{3} \end{aligned}$$

Of course you can also talk about conditional probabilities using random variables. Recall we have the random variable X:

$$X : S \to \{\text{Good}, \text{Bad}\}$$

where

$$\begin{aligned} X(\text{One}) = X(\text{Six}) &= \text{Good} \\ X(\text{Two}) = X(\text{Three}) = X(\text{Four}) = X(\text{Five}) &= \text{Bad} \end{aligned}$$

and the random variable Y:

$$Y : S \to \{\text{Even}, \text{Odd}\}$$

where

$$\begin{aligned} Y(\text{Two}) = Y(\text{Four}) = Y(\text{Six}) &= \text{Even} \\ Y(\text{One}) = Y(\text{Three}) = Y(\text{Five}) &= \text{Odd} \end{aligned}$$

Here's an example of conditional probabilities using random variables:

$$\Pr[X = \text{Good} \mid Y = \text{Even}]$$

This is

$$\Pr[X = \text{Good} \mid Y = \text{Even}] = \frac{\Pr[X = \text{Good and } Y = \text{Even}]}{\Pr[Y = \text{Even}]}$$

(look at the word "and"). Get it? The value is:

$$\begin{aligned}
\Pr[X = \textsc{Good} \mid Y = \textsc{Even}] &= \frac{\Pr[X = \textsc{Good} \text{ and } Y = \textsc{Even}]}{\Pr[Y = \textsc{Even}]} \\
&= \frac{p(\{s \in S \mid X(s) = \textsc{Good} \text{ and } Y(s) = \textsc{Even}\})}{p(\{s \in S \mid Y(s) = \textsc{Even}\})} \\
&= \frac{p(\{\textsc{Six}\})}{p(\{\textsc{Two}, \textsc{Four}, \textsc{Six}\})} \\
&= \frac{1/6}{1/6 + 1/6 + 1/6} \\
&= \frac{1}{3}
\end{aligned}$$

Exercise 9.6.2. Compute the following:

- $\Pr[X = \textsc{Good} \mid Y = \textsc{Odd}]$
- $\Pr[X = \textsc{Bad} \mid Y = \textsc{Even}]$
- $\Pr[X = \textsc{Bad} \mid Y = \textsc{Odd}]$
- $\Pr[Y = \textsc{Even} \mid X = \textsc{Good}]$
- $\Pr[Y = \textsc{Even} \mid X = \textsc{Bad}]$
- $\Pr[Y = \textsc{Odd} \mid X = \textsc{Good}]$
- $\Pr[Y = \textsc{Odd} \mid X = \textsc{Bad}]$

□

Exercise 9.6.3. Consider the random experiment of rolling two dice. The sample space is $S = S_1 \times S_2$ where $S_i = \{\textsc{One}, \textsc{Two}, \ldots, \textsc{Six}\}$ $(i = 0, 1)$. For instance $(\textsc{One}, \textsc{Two}) \in S$ is the outcome that the first die is One and the second die is Two. The probability is uniform. Let $X_1((s_1, s_2)) = s_1$ and $X_2((s_1, s_2)) = s_2$. Compute the following:

- $\Pr[X_1 = \textsc{One}]$
- $\Pr[X_2 = \textsc{Two}]$
- $\Pr[X_1 = \textsc{One} \mid X_1 \in \{\textsc{One}, \textsc{Two}\}]$

- $\Pr[X_1 = \text{One} \mid X_2 \in \{\text{One}, \text{Two}\}]$

□

Exercise 9.6.4. Consider the random experiment of rolling two dice. The sample space is $S = S_1 \times S_2$ where $S_i = \{\text{One}, \text{Two}, \ldots, \text{Six}\}$ $(i = 0, 1)$. For instance $(\text{One}, \text{Two}) \in S$ is the outcome that the first die is One and the second die is Two. The probability is uniform. Let $X_1((s_1, s_2)) = s_1$ and $X_2((s_1, s_2)) = s_2$.

- What is the probability of getting a sum of 6 if one of the die is less than 3? Before you compute that using intuition, I want you to write it down using a conditional probability, possibly with extra random variables.

- Now do the same for the probability of getting a sum of 3 if one of the die is less than 6.

□

File: discrete-probability/independence.tex

9.7 Independence

Suppose $p : S \to [0, 1]$ is a probability distribution function. Two events A and B with $p(A) > 0, p(B) > 0$ are **independent** if

independent

$$p(A \mid B) = p(A)$$

Recall that by definition

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

Therefore if A and B are independent, then

$$p(A) = p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

Therefore

$$p(A)p(B) = p(A \cap B)$$

Therefore the following are equivalent:

$$p(A \mid B) = p(A) \iff p(B \mid A) = p(B) \iff p(A \cap B) = p(A) \cdot p(B)$$

By the way $p(A \cap B)$ is called the **joint probability** of A and B.

joint probability

Intuitively, the fact that A and B are independent, i.e.,

$$p(A \mid B) = p(A)$$

means that the chance of A is not dependent on whether B has occurred or not.

Take for instance intuitively the probability of getting a one when you roll a die knowing that you get a one or two or three or four or five or six should be the same as getting a one. However, the probability of getting a one if I get a one or two is defintely higher than the probability of getting a one:

$$p(\{\text{One}\} \mid \{\text{One}, \text{Two}\}) = \frac{p(\{\text{One}\})}{p(\{\text{One}, \text{Two}\})} = \frac{1}{2} \neq p(\{\text{One}\})$$

In this case, B in $p(A \mid B)$ actually gives you more information.

You can also talk about the independence of two random variables:

Definition 9.7.1. Let X and Y be random variables on sample space S: $X : S \to V$ and $Y : S \to V'$ are functions. We say that X and Y are **independent** if

independent

$$\Pr[X = x \text{ and } Y = y] = \Pr[X = x] \cdot \Pr[Y = y]$$

for all $x \in X(S)$ and $y \in Y(S)$. The above condition is the same as

$$\Pr[X = x \mid Y = y] = \Pr[X = x]$$

which is the same as

$$\Pr[Y = y \mid X = x] = \Pr[Y = y]$$

Note that this definition does *not* depend on X and Y mapping to $\mathbb{R}$; in fact they don't even need to map to the same set. Note that $\Pr[X = x \text{ and } Y = y]$ means

$$\Pr[X = x \text{ and } Y = y] = p(\{s \in S \mid X(s) = x, Y(s) = y\})$$

I will also write $\Pr[X = x, Y = y]$ or $\Pr[(X = x) \wedge (Y = y)]$ for $\Pr[X = x \text{ and } Y = y]$.

Exercise 9.7.1. Let $S = \{\text{One}, \text{Two}, ..., \text{Six}\}^2$ (you are rolling two dice) and p be the uniform pdf on S. Let $X_1 : S \to \mathbb{R}$ be the random variable: $X_1(s,t) = 1$ if $s = t$ and 0 otherwise. Let $X_2 : S \to \mathbb{R}$ be the random variable: $X_2(s,t) = 1$ if $s \in \{\text{One}, \text{Two}, \text{Three}\}$ and 0 otherwise. Let $X_3 : S \to \mathbb{R}$ be the random variable: $X_3(s,t) = 1$ if $t \in \{\text{One}, \text{Three}, \text{Five}\}$ and 0 otherwise.

- Are X_1, X_2 independent?
- Are X_1, X_3 independent?
- Are X_2, X_3 independent?
- Are $X_1 + X_2, X_2 + X_3$ independent?

Theorem 9.7.1. *Let X and Y be independent. Then*

$$\mathrm{E}[XY] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$$

Proof. Exercise. □

9.7.1 An experiment involving two experiments

Let's consider the case of a random experiment that involves performing *two* random experiments, one after another. Consider a random experiment R that involves the tossing two fair coins, say I call the experiment of rolling the first coin R_1 and the second R_2. Suppose p_1 and p_2 be the pdf of the die 1 and die 2 respectively. For simplicity, suppose both coin are fair. I will denote the outcomes of R by

$$\begin{aligned} S &= \{\text{Head}, \text{Tail}\}^2 \\ &= \{(\text{Head}, \text{Head}), (\text{Head}, \text{Tail}), (\text{Tail}, \text{Head}), (\text{Tail}, \text{Tail})\} \end{aligned}$$

Since the two coins are fair, each outcome is equally likely:

$$p(x) = 1/4$$

for $x \in S$.

Consider the statement: "What is the probability of getting a tail for the second coin if the first coin is a head". We have two events. Let

$$A = \{\text{outcomes where the second toss gives a tail}\}$$

and

$$B = \{\text{outcomes where the first toss gives a head}\}$$

So "the probability of getting a tail for the second coin is the first coin is a head" which might be informally written as

$$p(\text{second coin} = \text{T} \mid \text{first coin} = \text{H})$$

Formally, of course A is just

$$A = \{(\text{Head}, \text{Tail}), (\text{Tail}, \text{Tail})\}$$

and

$$B = \{(\text{Head}, \text{Head}), (\text{Head}, \text{Tail})\}$$

Then

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)} = \frac{p(\{(\text{Head}, \text{Tail})\})}{p(\{(\text{Head}, \text{Head}), (\text{Head}, \text{Tail})\})} = \frac{1/4}{1/4 + 1/4} = 1/2$$

Here's another example. Suppose I roll a die and toss a coin. One would expect the output of the die to be independent of the coin. To be specific, the

event of getting a getting a six on the die to be independent of the event that we get a tail for the coin. Let's verify that. The sample space is

$$S = \{\text{Head}, \text{Tail}\} \times \{\text{One}, \text{Two}, ...\text{Six}\}$$

We want to verify that $p(A) = p(A \mid B)$ where $A = \{\text{Tail}\} \times \{\text{One}, ..., \text{Six}\}$ and $B = \{\text{Head}, \text{Tail}\} \times \{\text{Six}\}$. Then

$$\begin{aligned} p(A \mid B) &= \frac{p(A \cap B)}{p(B)} \\ &= \frac{p(\{(\text{Tail}, \text{Six})\})}{p(\{(\text{Head}, \text{Six}), (\text{Tail}, \text{Six})\})} \\ &= \frac{1/12}{2/12} \\ &= 1/2 \end{aligned}$$

and

$$\begin{aligned} p(A) &= p(\{(\text{Tail}, \text{One}), ..., (\text{Tail}, \text{Six})\}) \\ &= 6 \cdot \frac{1}{12} \\ &= 1/2 \end{aligned}$$

Exercise 9.7.2. Consider the random experiment of rolling a fair die. List all possible events A and B which are independent (if any). □

Exercise 9.7.3. Consider the random experiment of rolling two dice. The sample space is $S = S_1 \times S_2$ where $S_i = \{\text{One}, \text{Two}, \ldots, \text{Six}\}$ $(i = 0, 1)$. For instance $(\text{One}, \text{Two}) \in S$ is the outcome that the first die is One and the second die is Two. The probability is uniform. Let $X_1((s_1, s_2)) = s_1$ and $X_2((s_1, s_2)) = s_2$. Compute the following:

- Are the events $X_1 = \text{One}$ and $X_1 \in \{\text{One}, \text{Two}\}$ independent? "$X_1 = \text{One}$" is the event that the first die roll is one. "$X_1 \in \{\text{One}, \text{Two}\}$" is the event that the first die roll is either one or two.
- Are the events $X_1 = \text{One}$ and $X_2 \in \{\text{One}, \text{Two}\}$ independent?

The concept of independence for two events can be extended to a collection of any number of events. For the case when there are more than two events, there are two concepts of independence:

Definition 9.7.2. Let $A_1, A_2, \ldots, A_n$ be events.

1. $A_1, \ldots, A_n$ are **pairwise independent** if A_i and A_j independent for $i \neq j$, i.e., pairwise independent
$$p(A_i \cap A_j) = p(A_i)p(A_j)$$

2. $A_1, \ldots, A_n$ are **mutually independent** if for any collection of the above A_i's, say A_{i_0}, A_{i_1}, ... $A_{i_{k-1}}$ where $i_0 < i_1 < \cdots < i_{k-1}$, we have mutually independent
$$p(A_{i_0} \cap \cdots \cap A_{i_{k-1}}) = p(A_{i_0}) \cdots p(A_{i_{k-1}})$$

File: discrete-probability/sequence-of-experiment.tex

9.8 Sequence of Experiments

Frequently the random experiment you are interested in can be broken/decomposed into smaller and simpler experiments.

9.8.1 Sequence of independent experiments

One way where indicator random variables occur is when you need some kind of "if-else". For instance suppose you are playing a die-rolling gambling game and X_0 and X_1 are two "gain" functions. Why two? Because I'm changing the game so that you have to toss a coin. If your toss gives you a Head, your gain will be based on X_0. Otherwise your gain is based on X_1. Instead of doing two separate analysis based on the following two cases (based on the coin toss), you can combine them into one by defining two random variables:

$$X = I_{\text{Head}} \cdot X_0 + I_{\text{Tail}} \cdot X_1$$

where I_{Head} is the random variable

$$I_{\text{Head}}(\text{Head}) = 1, \;\; I_{\text{Head}}(\text{Tail}) = 0$$

and

$$I_{\text{Tail}}(\text{Head}) = 0, \;\; I_{\text{Tail}}(\text{Tail}) = 1$$

This means that in the game, where you made toss t (of a fair coin) and made roll r (of a fair die), if your toss t is a head, then you gain is given by

$$\begin{aligned} X(r) &= 1 \cdot X_0(r) + 0 \cdot X_1(r) \\ &= X_0(r) \end{aligned}$$

and if your toss is a tail, then your gain is

$$X(r) = X_1(r)$$

Here's a very important warning. Pay attention.

Your random experiment involves tossing a coin *and* rolling a die. The sample space should really be

$$S = S_0 \times S_1$$

where

$$S_0 = \{\text{Head}, \text{Tail}\}$$

and

$$S_1 = \{\text{One}, \text{Two}, \ldots, \text{Six}\}$$

Assuming the coin and die are both fair, we must have

$$p : S = S_0 \times S_1 \to [0, 1]$$
$$p(t, r) = 1/12$$

So even though I wrote

$$I_{\text{Head}}(\text{Head}) = 1, \;\; I_{\text{Head}}(\text{Tail}) = 0$$

i.e., I_{Head} is a random variable on

$$S_0 = \{\text{Head}, \text{Tail}\}$$

in the context of

$$X = I_{\text{Head}} \cdot X_0 + I_{\text{Tail}} \cdot X_1$$

I_{Head} is implicitly extended to

$$\widetilde{I}_{\text{Head}} : S_0 \times S_1$$

i.e.,

$$\widetilde{I}_{\text{Head}}(\text{Head}, r) = 1, \;\; \widetilde{I}_{\text{Head}}(\text{Tail}, r) = 0$$

for all rolls $r \in S_1$. In other words $\widetilde{I}_{\text{Head}}$ is the indicator variable $I_{\{\text{Head}\} \times S_1}$. In other words

$$\widetilde{I}_{\text{Head}}(\text{Head}, r) = I_{\text{Head}}(\text{Head}) = 1$$
$$\widetilde{I}_{\text{Head}}(\text{Tail}, r) = I_{\text{Head}}(\text{Tail}) = 0$$

for all rolls r. Furthermore the gain X_1

$$X_1 : S_1 \to \mathbb{R}$$

should really be extended to

$$\widetilde{X}_1 : S_0 \times S_1 \to \mathbb{R}$$

where

$$\widetilde{X}_1(t, r) = X_1(r)$$

So to be really accurate, we extend $I_{\text{Head}}, I_{\text{Tail}}, S_0, S_1$ to $\widetilde{I}_{\text{Head}}, \widetilde{I}_{\text{Tail}}, \widetilde{S}_0, \widetilde{S}_1$ and then define X (in the correct way) as

$$X = \widetilde{I}_{\text{Head}} \cdot \widetilde{X}_0 + \widetilde{I}_{\text{Tail}} \cdot \widetilde{X}_1$$

i.e.,

$$X(t,r) = \widetilde{I}_{\text{Head}}(t,r) \cdot \widetilde{X}_0(t,r) + \widetilde{I}_{\text{Tail}}(t,r) \cdot \widetilde{X}_1(t,r)$$

and then we do get

$$X(t,r) = I_{\text{Head}}(t)X_0(r) + I_{\text{Tail}}(t)X_1(r)$$

By the way, the events "the toss is head" and "the die roll is one" are independent. In other words

$$\begin{aligned} A &= \{\text{Head}\} \times S_1 \\ B &= S_0 \times \{\text{One}\} \end{aligned}$$

are independent. Here's the check:

$$\begin{aligned} p(A) &= p(\{\{\text{Head}\} \times S_1\}) = 6/12 \\ p(B) &= p(S_0 \times \{\text{One}\}) = 2/12 \\ \therefore \quad p(A) \cdot p(B) &= 1/12 \end{aligned}$$

and

$$p(A \cap B) = p(\{(\text{Head}, \text{One})\}) = 1/12$$

Hence

$$p(A \cap B) = p(A) \cdot p(B)$$

In fact A and B are independent even when A is replaced by "any toss" and B is replaced by "any roll".

Another thing to note is that

$$p(t,r) = p_0(t) \cdot p_1(r)$$

where $p_0 : \{\text{Head}, \text{Tail}\} \to [0,1]$ is the uniform pdf of a (fair) coin toss and $p_1 : \{\text{One}, \text{Two}, \text{Three}, ..., \text{Six}\} \to [0,1]$ is the uniform pdf of a (fair) die roll.

Exercise 9.8.1. Suppose in a book, you read: "Consider a random experiment of rolling two fair dice. The probability that the first die is even is 1/2 and the

probability that the second die is odd (i.e., one, three, or five) is 1/2. Therefore the probability that the first is even and the second is odd is $(1/2) \cdot (1/2) = 1/4$." Let $p_0 : S_0 \to [0, 1]$ denote the pdf of the random experiment of rolling only the first die. Let $p_1 : S_1 \to [0, 1]$ denote the pdf of the random experiment of rolling only the second die. Let $p : S \to [0, 1]$ be the pdf of the pdf of random experiment of rolling the first and then the second die.

- What is the sample space S?
- What is the value $p(r_0, r_1)$?
- What is the event A of "the first roll is even" in the random experiment of rolling twi dice?
- What is the event B of "the second roll is odd" in the random experiment of rolling twi dice?
- Prove formally that A and B are independent.
- What is the value p in terms of p_0 and p_1? □

Theorem 9.8.1. *If a random experiment involves carrying out two separate and independent experiments. Suppose the pdfs are* $p_0 : S_0 \to [0, 1]$ *and* $p_1 : S_1 \to [0, 1]$. *Then*

$$p(\{x\} \times S_2) = p_1(x)$$

and

$$p(S_1 \times \{y\}) = p_2(x)$$

and

$$p(x, y) = p_1(x)p_2(y)$$

This means that if there is a random experiment (with pdf $p : S \to [0, 1]$) which is made up of a sequence of two independent experiment (with pdf $p_i : S_i \to [0, 1]$), then $p(x, y) = p_0(x)p_1(y)$, which means that the computation simplifies to the computation of $p_0(x)$ and $p_1(y)$ and note that the computation of $p_0(x)$ for instance allows you to focus on $p_0 : S_0 \to [0, 1]$ and this random experiment has a smaller sample space and is therefore simpler.

The above is also true for a sequence of two independent three random experiments, or four, or five, etc.

9.8.2 Sequence of dependent experiments

There are times when a random experiment involves two experiments and the second experiment depends on the first. Here's an example. Suppose you have two boxes: The first box has 1 red ball and 2 green balls while the second box has 3 red balls and 5 green balls. Here's the experiment: I toss a fair coin. If I get a head, then I pick a ball from the first box. If I get a tail, then I pick a ball from the second box.

It's clear that the sample space is

$$S = \{\textsc{Head}, \textsc{Tail}\} \times \{\textsc{Red}, \textsc{Green}\}$$

The pdf p of the above random experiment is then

$$\begin{aligned} p(\textsc{Head}, \textsc{Red}) &= 1/2 \cdot 1/3 = 1/6 \\ p(\textsc{Head}, \textsc{Green}) &= 1/2 \cdot 2/3 = 1/3 \\ p(\textsc{Tail}, \textsc{Red}) &= 1/2 \cdot 3/8 = 3/16 \\ p(\textsc{Head}, \textsc{Green}) &= 1/2 \cdot 5/8 = 5/16 \end{aligned}$$

Is the event that I get a head (from the coin toss) independent of the event that I get a red ball?

Let A be the first event. This means

$$A = \{(\textsc{Head}, \textsc{Red}), (\textsc{Head}, \textsc{Green})\}$$

Let B be the second event. Then

$$B = \{(\textsc{Head}, \textsc{Red}), (\textsc{Tail}, \textsc{Red})\}$$

We have

$$\begin{aligned} p(A \mid B) &= \frac{p(A \cap B)}{p(B)} \\ &= \frac{p(\{(\textsc{Head}, \textsc{Red})\})}{p(\{(\textsc{Head}, \textsc{Red}), (\textsc{Tail}, \textsc{Red})\})} \\ &= \frac{1/6}{1/6 + 3/16} \\ &= \frac{8}{17} \end{aligned}$$

which is not $p(A) = 1/6$. Therefore A and B are not independent.

9.8.3 Multiplication principle

Suppose that a random experiment R involves carry out two experiments. After getting an outcome x from the first random experiment R_0. Say the pdf of R_0 is $p_0 : S_0 \to [0, 1]$. I perform the second random experiment R_1. The second random experiment might depend on the first, i.e., I should write $R_1(x)$ – what I'm going to say deos not depend on whether the two experiments are dependent or not. Suppose the pdf for the second experiment is $p_{1,x} : S_{1,x} \to [0, 1]$. (Note that it's possible that $p_{1,x}$ depends on the first outcome x.) Suppose the output for the second random experiment is y. Then the probability for outcome (x, y) is $p_0(x) \cdot p_{1,x}(y)$.

For instance if you roll two dice where the first is fair and the second is loaded so that 1/2 the time you will get a six, then the probability of getting one for the first and six for the second is

$$1/6 \cdot 1/2$$

Or, if you toss one fair coin and if you get a head, you roll a fair die and if you get a tail, you roll a loaded die where the chance of one is 0.5 and the others are equi-distributed. Then the probability of head followed by one is (1/2)(1/6) and the probability of tail followed by one is (1/2)(1/2).

The above fact follows from the multiplication principle of counting.

9.8.4 Variable length

There are times when you are interested in an experiment that is made up of a sequence of experiments where the number of experiments n is not fixed. In fact the questions you are interested in involves solving for n when a condition is met.

For instance: How times do I have a die in order so that the chance of getting two consecutive sixes is greater than 1/2?

File: discrete-probability/birthday.tex

9.9 The Birthday Paradox

How many people do you need in order to have at least a pair with the same birthday? There are about 365.25 days in a day. So you would think that you need maybe about half. Of course the pigeonhole guarantees there will be such a pair if you have 366 people.

But surprisingly, you actually need a very small group of randomly chosen people in order to have a good chance (i.e., probability of > 0.5) that at least one pair has a common birthday.

Since I'm trying to prove that a very small number is enough, let's round up and say that there are 366 days in a year.

It's easier to ask the opposite question first. Let n be the number of people you have chosen. What is the probability that there is *no* two with the same birthday?

That sounds kind of difficult. So let's make things concrete, What if $n = 1$? Well ... there's no way for a pair to have the same birthday! Because there's only one person! Duh!

OK. What if there are two? In other words, what if $n = 2$? Suppose the birthday of the first person is known. Then in order for the second person not to have the same birthday as the first, the second person has a "choice" of birthday among $366 - 1$ days out of 366 days. So in this case, the probability is

$$\frac{366 - 1}{366}$$

Now what if $n = 3$? Suppose there are 3 people. After the first person has announced his/her birthday, then the probability that the second person has a different birthday as the first must be

$$\frac{366 - 1}{366}$$

After that, since two birthdays are taken, the probability for the third person to have a birthday different from the first two must be

$$\frac{366 - 2}{366}$$

Altogether, the probability that among 3 randomly chosen people not to have the same birthday must be

$$\frac{366-1}{366} \cdot \frac{366-2}{366}$$

In general, you see quickly that if you have n people, the probability that they all have different birthdays must be

$$\frac{366-1}{366} \cdot \frac{366-2}{366} \cdot \cdots \cdot \frac{366-(n-1)}{366}$$

Of course the probability of have at least a pair with the same birthday is then

$$1 - \frac{366-1}{366} \cdot \frac{366-2}{366} \cdot \cdots \cdot \frac{366-(n-1)}{366}$$

Now I'm interested in knowing when there will be at least two with the same birthday. Probabilistically speaking, that means I want the probability is greater than 0.5. So I want to find n such that

$$1 - \frac{366-1}{366} \cdot \frac{366-2}{366} \cdot \cdots \cdot \frac{366-(n-1)}{366} > 0.5$$

We can find n by running the following program:

```
def f(i):
    return (366 - i) / 366.0

def prob(n):
    p = 1
    for i in range(1, n):
        p *= f(i)
    return 1 - p

for n in range(1, 40):
    print (n, prob(n))
```

Here's the output:

```
1 0
2 0.00273224043716
3 0.00818179103586
4 0.0163114484864
5 0.0270621430384
6 0.0403536438166
```

```
7 0.056085551295
8 0.0741385598768
9 0.0943759684041
10 0.116645411804
11 0.140780783066
12 0.166604311444
13 0.193928760249
14 0.222559705923
15 0.252297859249
16 0.282941389607
17 0.314288214105
18 0.34613821509
19 0.378295352052
20 0.410569637055
21 0.442778946506
22 0.474750646296
23 0.506323011819
24 0.537346429109
25 0.567684368184
26 0.597214124456
27 0.625827328729
28 0.653430230708
29 0.679943764971
30 0.705303412009
31 0.729458870041
32 0.752373555912
33 0.774023955395
34 0.794398844663
35 0.813498405541
36 0.83133325747
37 0.847923428867
38 0.863297289883
39 0.877490467436
```

Looks like we just need 23, which is surprising small!

To verify this, just get a group of 23 people in a room and ask for their birthdays then there's a strong likelihood that there are two with the same birthdays.

It's troublesome to get 23 people together. So let's write a simple program to pick 23 numbers from 1 to 366 and see if there are two with the same value.

```
import random; random.seed()
```

```
xs = [random.randrange(1, 367) for _ in range(23)]
xs.sort()
print xs
```

Here's the output running the above several times. I've done it 7 times. Go ahead and count the number of times where there is a repeat birthday:

```
[28, 57, 92, 101, 114, 134, 138, 174, 176, 179, 202, 208,
218, 231, 239, 241, 251, 253, 283, 286, 294, 317, 325]
[19, 40, 50, 53, 64, 74, 129, 140, 146, 161, 169, 179,
193, 194, 221, 231, 249, 269, 271, 293, 321, 350, 357]
[14, 36, 43, 55, 83, 124, 133, 133, 138, 155, 256, 263,
267, 274, 283, 296, 307, 315, 325, 338, 348, 349, 355]
[4, 41, 44, 80, 119, 124, 136, 139, 140, 163, 183, 212,
215, 216, 217, 228, 235, 248, 256, 263, 271, 281, 350]
[25, 28, 67, 91, 120, 146, 152, 152, 196, 206, 222, 223,
223, 234, 240, 242, 248, 253, 288, 289, 314, 320, 352]
[20, 21, 55, 76, 91, 102, 104, 119, 144, 152, 173, 200,
204, 206, 214, 222, 230, 249, 249, 326, 334, 354, 360]
[24, 30, 35, 50, 62, 66, 77, 78, 91, 91, 121, 150, 153,
183, 184, 211, 211, 220, 235, 262, 344, 347, 351]
```

Now I'm going to do this 10000 times:

```
def repeat(xs):
    for i in range(22):
        if xs[i] == xs[i + 1]:
            return True
    return False

import random; random.seed()
count = 0
for i in range(10000):
    xs = [random.randrange(1, 367) for _ in range(23)]
    xs.sort()
    if repeat(xs):
        count += 1

print count
```

Here's my output:

```
5103
```

So 5103 out of 10000 simulations have a repeat birthday. When I ran the experiment with 100000 simulations, the number is

```
50519
```

In all cases, more than half of the cases of randomly generated 23 birthdays has at least two with the same birthday.

Instead of using a program, let's look at the above expression again:

$$1 - \frac{366-1}{366} \cdot \frac{366-2}{366} \cdot \cdots \cdot \frac{366-(n-1)}{366} > 0.5$$

This is

$$1 - \left(1 - \frac{1}{366}\right)\left(1 - \frac{2}{366}\right) \cdots \left(1 - \frac{n-1}{366}\right) > 0.5$$

We want to find the smallest n satisfying the above. You can tell that this is not one of the standard equations from your math classes. So let me solve this inequality using a program. (An approximation using math is below.) Here's the program:

```
n = 2
while 1:
    p = 1
    for i in range(1, n):
        p *= (1 - i / 366.0)
    print (n, 1 - p)
    if 1 - p > 0.5:
         break
    n += 1
```

and the output

```
[student@localhost book] python a15245236.py
2 0.002732240437158473
3 0.008181791035862584
4 0.016311448486388325
5 0.02706214303844967
6 0.040353643816612994
7 0.056085551295029235
8 0.07413855987681828
9 0.09437596840410101
10 0.11664541180400023
11 0.14078078306618602
12 0.16660431144397825
13 0.19392876024909378
14 0.22255970592330632
15 0.25229785924864434
16 0.2829413896073065
17 0.3142882141053478
18 0.3461382150895257
19 0.3782953520523359
20 0.4105696370550834
21 0.4427789465056253
22 0.47475064629628616
23 0.5063230118194602
```

Now let's derive that using math (and a calculator). Note that

$$\begin{aligned}\left(1-\frac{1}{366}\right)\left(1-\frac{2}{366}\right)\cdots\left(1-\frac{n-1}{366}\right) &\leq e^{-1/366}e^{-2/366}\cdots e^{-(n-1)/366}\\ &= e^{-(1+2+\cdots+(n-1))/366}\\ &= e^{-\frac{(n-1)n}{2}/366}\end{aligned}$$

I'm using the inequality

$$1 + x \leq e^x$$

which comes from

$$1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots = e^x$$

if you recall the Maclaurin/Taylor series for e^x. Or you can derive this by proving $e^x - 1 - x \geq 0$ (standard Calc 1 type problem: show the left-hand side is ≥ 0 when $x = 0$ and the slope is > 0 for $x > 0$.)

Therefore the above inequality becomes

$$1 - \left(1 - \frac{1}{366}\right)\left(1 - \frac{2}{366}\right)\cdots\left(1 - \frac{n}{366}\right) \geq 1 - e^{-\frac{(n-1)n}{2}/366}$$

In that case for an n such that $1 - \left(1 - \frac{1}{366}\right)\left(1 - \frac{2}{366}\right)\cdots\left(1 - \frac{n-1}{366}\right)$ is approximately 0.5, we have

$$0.5 \approx 1 - e^{-\frac{(n-1)n}{2}/366}$$

which gives us

$$e^{-\frac{(n-1)n}{2}/366} \approx 0.5$$

Taking natural logs,

$$(n-1)n \approx (-\ln 0.5) \cdot 738$$

and therefore

$$n^2 - n + \ln 0.5 \cdot 738 \approx 0$$

Solving the quadratic we get

$$n \approx \frac{1 + \sqrt{1 - 4\ln 0.5 \cdot 738}}{2} = 23.1228\ldots$$

(of course taking the positive root).

If instead of 366 days in a year, if there are N days in a year (say we're on a different planet), then the minimum value of n to reach about 0.5 is given by

$$n \approx \frac{1 + \sqrt{1 - 4\ln 0.5 \cdot (2N)}}{2} = \frac{1 + \sqrt{1 - 8N\ln 0.5}}{2}$$

taking the positive root of the quadratic equation.

Exercise 9.9.1. Go through the above computation and tell me what is the approximate n so that the probability of finding a repeat birthday in randomly chosen n randomly chosen people is about 0.9. Write a program and verify. □

Exercise 9.9.2. What is the smallest n such that the probability of finding two birthdays among n people which are either on the same day or one day apart. □

Exercise 9.9.3. What is the smallest n such that in a group of n randomly chosen people, there are least 3 with the same birthdays. □

The birthday paradox is equivalent to the following question: Let there be n people. They all have been randomly given a value v from a set of size k in the following way: Your k value are written on a piece of paper, you close your eyes and point your finger, and that value is given to a person. You repeat until every person in your group of n people has been assigned a value. What is the probability that at least two have been given the same value?

The birthday paradox is important in cryptography and hash tables because of hash collisions.

In the case of the hash table data structure, if you choose an array of linked list as implementation, collisions imply a longer linked list for each hash value of the keys and the worse runtime is determined by the length of the longest linked list.

In the case of digital signatures, assuming you are using the RSA version with pair of keys (k_0, k_1), to sign a message m, Alice would compute $m' = D(h(m), k_1)$ which would be the digital signature. Here h is the hash function and $(E(k_0, \bullet), D(k_1, \bullet))$ are the encryption-decryption pair of functions for Alice's keys (k_0, k_1) where k_0 public and k_1 is private. Alice sends $(m, m', h(m))$ to Bob. Bob can read the message m, but Bob also want to be certain that m is from Alice. So Bob computes

$$E(k_0, m')$$

If $m' = D(k_1, h(m))$, i.e., it's really created by Alice (assuming no one else has k_1, only Alice can create m'), then $E(k_0, m')$ must be $h(m)$.

Suppose Eve intercepts $(m, m', h(m))$. She then creates another message m^* such that $h(m^*) = h(m)$ and then sends $(m^*, m', h(m))$ to Bob. Bob computes

$$E(k_0, m') = h(m) = h(m^*)$$

and thinks that m^* is signed by Alice. Bob is counting on the fact that it's difficult to find another m^* such that $h(m^*) = h(m)$. (The above is a textbook example of digital signatures. In real life, there are a few more things to do. Furthermore the message m itself is usually encrypted.)

File: discrete-probability/average-examples.tex

9.10 Examples of average computation

Now I'm going to do some very important examples on the computation of average values. You'll see the frequent use of indicator random variables. Recall this exercise from the section on indicator random variable: Let X_A be an indicator random variable on $A \subseteq S$. Then

$$\mathrm{E}[X_A] = p(A)$$

See Exercise 9.5.1 (on page 8035).

Example 9.10.1. What is the sum of rolling two fair dice?

Of course you can't answer that exactly. When you read something like the above question, you should immediately replace the question by "What is the *average* of the sum of rolling two fair dice?"

Do this slowly – don't rush. First set up the notation of the problem. Let $p : S \to [0, 1]$ be the pdf of this random experiment. The random experiment is "roll two dice". The outcomes are

$$S = \{(\textsc{One}, \textsc{One}), (\textsc{One}, \textsc{Two}), ..., (\textsc{Six}, \textsc{Six})\}$$

and p is uniform, i.e., $p(s) = 1/36$. The random variable is $X : S \to \mathbb{R}$ where

$$\begin{aligned} X(\textsc{One}, \textsc{One}) &= 2 \\ X(\textsc{One}, \textsc{Two}) &= 3 \\ &\vdots \\ X(\textsc{Six}, \textsc{Six}) &= 12 \end{aligned}$$

Therefore the answer is

$$\begin{aligned} \mathrm{E}[X] &= \sum_{s \in S} X(s) \cdot p(s) \\ &= \sum_{x \in X(S)} x \cdot \Pr[X = x] \\ &= 2 \cdot (1/36) + 3 \cdot (2/36) + \cdots + 12 \cdot (1/36) \end{aligned}$$

You can check that it's 7.

But ... there's another way to do this ...

Let X_i be the random variable of number of dots on the top of die i when you roll it. Let $X = X_1 + X_2$. Make sure you write down the complete description of X_i and the associated probability $\Pr[X_i = \bullet]$. For instance

$$X_1(\text{One}, \text{Three}) = 1$$

and

$$X_2(\text{One}, \text{Three}) = 3$$

and

$$X(\text{One}, \text{Three}) = X_1(\text{One}, \text{Three}) + X_2(\text{One}, \text{Three}) = 1 + 3$$

Etc. Do NOT skip the proper definitions of your sample space, the random variables, etc. As I mentioned in the previous section, books like to pretend, while working on X_1 that there's only one die and write

$$X_1(\text{One}) = 1$$

You should really think of X_1 as a function on the full sample space first.

Then

$$\mathrm{E}[X_i] = 1(1/6) + 2(1/6) + \cdots + 6(1/6)$$

which you can check is $7/2$. Therefore by the linearity of the sum of random variables,

$$\mathrm{E}[X] = \mathrm{E}[X_1 + X_2] = \mathrm{E}[X_1] + \mathrm{E}[X_2] = 7$$

Neat right? Make sure you read the above two solutions *very* carefully and understand the benefit of the linearity of the sum of random variables. □

Exercise 9.10.1. What is the sum of even rolls (i.e., don't count odds) if you roll n fair dice? Compute it by hand and then write a program to verify. □

Exercise 9.10.2. You have m n–sided fair dice. You roll die #1. If your roll #1 is 3, you get to roll the second die 3 times (or you can choose to roll it just once or twice). Suppose you roll #2 twice and you get 4. That means you can roll your third die 4 times. Etc. The goal is to maximize the sum of your rolls. You have decided that the strategy is to stop rolling a die if it gives you a 4 or 5 or 6. What is the sum? □

Example 9.10.2. Card Shuffling Problem. Suppose you have a set of n (distinct) cards. For simplicity suppose the cards are numbered $0, 1, 2, ..., n-1$. Suppose you then give them a good shuffle so that the desk of cards is now random, i.e., each card is given a random position. Here's the question: How many cards are in their original position before the shuffle?

Card Shuffling Problem

This problem is the same as the Hat Check Problem: If n men arrive at a restaurant, leave their hats at the reception, and collect their hats on the way out, assuming an evil receptionist who gave the hats back randomly – and the n men are half drunk and did not look carefully at their hats – how many men will receive their own hats?

Hat Check Problem

This is also called the Envelope Problem where an incompetent secretary randomly puts letters into addressed envelopes.

Envelope Problem

Of course you can't give me an answer that works for all cases because the cases are random. You should interpret the above question as: "*On the average*, how many cards are in their original position before the shuffle?" What this means is that if you claim that 5 cards retain their original position, then you are actually saying that if you perform the above experiment 1,000,000,000 times, then average across all these 1,000,000,000,000 cases, the number of cards retaining their original position is 5.

This problem can be very difficult because there are too many cases to consider. You would for instance have to consider the case where card 0 is at position 0 and card 0 not at position 0. Then when consider card 1, you want to worry about the options for card 1 wrt card 0. For instance if card 0 is at position 0, then card 1 don't have the option of being shuffle to position 0. But in the case when card 1 is shuffled to position 1, then card 1 cannot be at position 1. Etc. YIKES.

Here's another way to solve this problem.

Formally, the sample space S is the set of all permutations of $(0, 1, 2, ..., n-1)$. Imagine the cards are labeled $0, 1, 2, 3, ..., n-1$ before being shuffled. For the case when $n = 4$, say after a shuffle, the outcome is $(2, 0, 1, 3)$. This represents the deck of cards where: (card numbered 2, card numbered 0, card numbered 1, card numbered 3). In other words, the outcomes in S represents all possible shuffled.

The pdf $p : S \to [0, 1]$ is uniform since shuffle will result in a permutation of the cards without any preference for any particular permutation.

Define X_i to be the random variable so that $X_i = 1$ if the i–th card is at the i–position after the shuffle. Then $X_i(s_0, s_1, s_2, \ldots, s_{n-1}) = 1$ if $s_i = i$ if $(s_0, s_1, s_2, \ldots, s_{n-1}) \in S$. Note that there are n random variables X_i ($i = 0, 1, ..., n-1$). Another thing to note is that X_i is an indicator random variable on the event $A_i = \{(s_0, s_1, s_2, ..., s_{n-1}) \in S \mid s_i = i\}$.

Now define another random variable

$$X = \sum_{i=0}^{n-1} X_i$$

This is clearly a random variable on S. Also, note that $X(s) = \sum_{i=0}^{n-1} X_i(s)$ is the number of $i \in \{0, 1, 2, ..., n-1\}$ such that $s_i = i$. See it? Therefore the required number, the average number of cards which retained their original position, is $\mathrm{E}[X]$. Therefore

$$\mathrm{E}[X] = \mathrm{E}\left[\sum_{i=0}^{n-1} X_i\right] = \sum_{i=0}^{n-1} \mathrm{E}\left[X_i\right]$$

The key thing is this: In the above case-by-case (incomplete analysis ... look for the word YIKES) of the cards, when I consider card 1, I have to worry about where card 1 can go to based on card 0. However in the computation of the expected value of a sum, I can compute the expected value of the terms separately even when the random variables are not independent – *this is key* to understanding the power of the linearity of the expectation of the sum of random variables.

Now what is $\mathrm{E}[X_i]$? Well that's

$$\mathrm{E}[X_i] = \sum_{s \in S} X_i(s) p(s)$$

where $p : S \to [0, 1]$ is the uniform pdf on S i.e., $p(s) = 1/|S| = 1/n!$.

$$\mathrm{E}[X_i] = \frac{1}{n!} \sum_{s \in S} X_i(s)$$

Now $\sum_{s \in S} X_i(s)$ is the number of permutations of S where i is fixed. Therefore $\sum_{s \in S} X_i(s) = (n-1)!$. Hence

$$\mathrm{E}[X_i] = \frac{1}{n!} \sum_{s \in S} X_i(s) \frac{1}{n!} \cdot (n-1)! = \frac{1}{n}$$

Hence

$$\mathrm{E}[X] = \sum_{i=0}^{n-1} E\left[X_i\right] = \sum_{i=0}^{n-1} \frac{1}{n} = 1$$

Which is interesting ... because the average is 1, which does not depend on the number of cards.

Note that I think of the X_i as a random variable on S (the full sample space). You can also restrict that to just the "sample space of the i–th card". In that case $X_i : \{0, 1, 2, ..., n-1\} \to \mathbb{R}$ where $X_i(s_i) = 1$ if $s_i = i$, otherwise it's 0. The corresponding pdf on $p_i : \{0, 1, 2, ..., n-1\} \to [0, 1]$ is $p_i(s_i) = 1/n$. You will arrive at the same result.

Exercise 9.10.3. Write a program to simulation card shuffling for different n cards and check that the average number of cards retained its original position if 1. □

Exercise 9.10.4. How many cards are "consecutives". For instance when $n = 4$, the deck $(1, 3, 2, 4)$ has one consecutive 3,2, i.e., adjacent cards with consecutive numbers. The deck $(1, 2, 4, 3)$ has two consecutives. □

Exercise 9.10.5. n men go back to a restaurant and check in their hats and coats. The evil receptionist gave them their hat and coat randomly. The men were drunk and did not check what they received. How many men will will get either their hats or their coats back? (The "or" is not exclusive.) □

Exercise 9.10.6. n men go back to a restaurant. Half of them were wearing hats. Those with hats check in their hats. On their way out, the evil receptionist gave them their hat randomly. The men were drunk and did not check what they received. Note that some of them who did not come with a hat might have received a hat (yes, they were that drunk). How many men received their own hats back? How many men who did not have hats received a hat? How many men who came with a hat did not receive a hat? □

Exercise 9.10.7. $2n$ men go to the same restaurant, but these men went in as pairs of friends: the first two are friends, the next two are friends, etc. The first pair are not friends of other men. When they go home, they receive their hats randomly. But they are drunk and did no check their hats. When they wake up tomorrow morning, if they did not receive their own hats, they can still check with their friend and see if the situation can be improved by exchanging their hats, and be happy. How many men are happy? □

Example 9.10.3. Inversion Problem. What is the average number of (pairs of) inversions in a randomly selected permutation of $\{0, 1, 2, ..., n-1\}$? For instance, for $(1, 3, 2)$, there is 1 pair of inversion. for $(1, 3, 2, 5, 4)$, there are 2 pairs of inversion. This one is more complicated: there are eight inversions in $(5, 3, 2, 4, 1)$ – see them? □

Inversion Problem

Let S be the sample space of permutations of $0, 1, 2, ..., n-1$:

$$S = \{(s_0, s_1, \ldots, s_{n-1}) \mid (s_0, s_1, \ldots, s_{n-1}) \text{ is a permutation of } (1, 2, 3, ..., n-1)\}$$

Let X_{ij} be a random variable such that $X_{ij}((s_0, ..., s_{n-1}))$ is 1 if s_i, s_j is an inversion, i.e., $s_i > s_j$. Otherwise X_{ij} is 0. Therefore the random variable

$$X = \sum_{0 \le i < j \le n-1} X_{ij}$$

counts the number of inversion. Therefore the required number is

$$\mathrm{E}[X] = \sum_{0 \le i < j \le n-1} \mathrm{E}[X_{ij}]$$

For $0 \le i < j \le n-1$,

$$\mathrm{E}[X_{ij}] = \frac{|\{(s_0, ..., s_{n-1}) \in S \mid s_i > s_j\}|}{|S|}$$

Note that we have the following disjoint union:

$$S = \{(s_0, ..., s_{n-1}) \in S \mid s_i > s_j\} \dot\cup \{(s_0, ..., s_{n-1}) \in S \mid s_i < s_j\}$$

(the bijection between the two partitions is just swapping the i– and j–coordinates). Therefore

$$\mathrm{E}[X_{ij}] = 1/2$$

and hence

$$\mathrm{E}[X] = \binom{n}{2} \cdot \frac{1}{2} = \frac{n(n-1)}{4}$$

□

Exercise 9.10.8. You should write a program to verify the above computation.

Exercise 9.10.9. There is a buckets of red, green, blue balls (R,G,B). The ratio R:G:B is 1:2:3. You close your eyes and pick n balls and put them on a straight line. What is the number of consecutive R-G-B triples on the straight line? □

Exercise 9.10.10. Continuing the example above, how many neighboring inversions are there? A **neighboring inversion** is an inversion next to each other. For example if you look at the permutation $(4, 2, 1, 3)$, you see the $4, 2$ is a neighboring inversion because they are next to each other in the permutation

neighboring inversion

$$(\underline{4, 2}, 1, 3)$$

(I'm talking about the index positions, not the value). However $4, 1$ is an inversion but not a neighboring inversion:

$$(\underline{4}, 2, \underline{1}, 3)$$

Write a program to test your conclusion. □

Exercise 9.10.11. m balls are randomly thrown into n buckets.

(a) How many balls are there in each bucket?

(b) How many buckets are empty?

(c) How many buckets contain more than 1 ball?

(d) How many buckets contain exactly 2 balls?

SOLUTION.

(a) Let $X_{ij} = 1$ if ball i falls into bucket j; otherwise it's 0. Then $X_j = \sum_i X_{ij}$ is the number of balls in bucket j. Therefore the number of balls in bucket j is $\mathrm{E}[X_j] = \sum_i \mathrm{E}[X_{ij}] = \sum_i (1/n) = m/n$.

(b) Let $X_i = 1$ if bucket i is empty. Let $X_{ij} = 1$ if ball j does not fall into bucket i. Therefore

$$X_i = \prod_j X_{ij}$$

Let $X = \sum_i X_i$. This is the number of buckets which are empty. and hence

$$\begin{aligned}
\mathrm{E}[X] &= \mathrm{E}\left[\sum_i X_i\right] \\
&= \mathrm{E}\left[\sum_i \prod_j X_{ij}\right] \\
&= \sum_i \mathrm{E}\left[\prod_j X_{ij}\right] \\
&= \sum_i \prod_j \mathrm{E}\left[X_{ij}\right] \\
&= \sum_i \prod_j (1 - 1/n) \\
&= \sum_i (1 - 1/n)^m \\
&= n(1 - 1/n)^m
\end{aligned}$$

(I used the fact that $\mathrm{E}[\prod_j X_{ij}] = \prod_j \mathrm{E}[X_{ij}]$ if X_{ij} (all j) are independent.)

By the way

$$\lim_{n\to\infty} (1 + 1/n)^n = e$$

Therefore

$$\lim_{n\to\infty} (1 - 1/n)^n = e^{-1}$$

Therefore for large n (number of buckets)), $n(1 - 1/n)^m = ne^{-m/n}$. And if $m = n$, $n(1 - 1/n)^m = ne^{-m/n} = n/e$ which means that $1/e$ of the buckets are empty. Note that $1/e$ is independent of n.

Alternative. Let $X_i = 1$ if the bucket i is empty. Then

$$\mathrm{E}[X_i] = \left(\frac{n-1}{n}\right)^m = \left(1 - \frac{1}{n}\right)^m$$

Let $X = \sum_i X_i$. Hence

$$\mathrm{E}[X] = \sum_i \mathrm{E}[X_i] = n\left(\frac{n-1}{n}\right)^m$$

which is the same as above.

Example 9.10.4. Coupon Collector Problem. There are n types of coupons for free coffee, one for each of the n different types of coffee at Starbucks. One coupon is randomly chosen and placed in identical Kellogg's cereal boxes. Since you are a coffee lover, you want to collect all the different n types of coupon. (Don't worry. Walmart has a very large number of such Kellogg's cereal boxes with Starbucks coffee coupon and they promise you that they do have all the n types of coupons are in their Kellogg's cereal boxes.) How many boxes of cereal do you need to buy to have a complete collection of all the n types of coupons?

Coupon Collector Problem

SOLUTION.

First I buy one box. This means that I have 1 type of coupon. Let X_1 be the number of boxes I need to buy to get 1 type of coupon. Clearly $\mathrm{E}[X_1] = 1$. (In fact $X_1 = 1$ is a constant random variable.)

Next, I buy enough boxes to get the second type of coupons. Let X_2 be the number of these boxes to get the second type of coupon. Suppose the sequence of coupons purchase (after the first box) is

$$(s_0, s_1, s_2, ...)$$

The sequence might be

$$(0, 0, 0, \text{not } 0, ...)$$

For this case, $X_2(0, 0, 0, \text{not } 0, ...)) = 4$. Of course you might be lucky: $X_2(\text{not } 0, ...)) = 1$. When I look at s_0, the probability that it is not 0 is $(n-1)/n$. Suppose s_0 is in fact 0. Then the probability that $s_1 \neq 0$ is $(n-1)/n$. And the probability that $s_0 = 0$ and $s_1 \neq 0$ is $(1/n)((n-1)/n)$. The probability that s_2 is the first coupon that is not zero is $(1/n)^2((n-1)/n)$. The probability that s_3 is the

first coupon that is not zero is $(1/n)^3((n-1)/n)$.

$$\begin{aligned}
\mathrm{E}[X_2] &= 1 \cdot \frac{n-1}{n} + 2 \cdot \left(\frac{1}{n}\right) \frac{n-1}{n} + 3 \cdot \left(\frac{1}{n}\right)^2 \frac{n-1}{n} + \cdots \\
&= \left[1 + 2 \cdot \left(\frac{1}{n}\right) + 3 \cdot \left(\frac{1}{n}\right)^2 + \cdots\right] \left(1 - \frac{1}{n}\right) \\
&= \left[1 + 2 \cdot \left(\frac{1}{n}\right) + 3 \cdot \left(\frac{1}{n}\right)^2 + \cdots\right] - \left[\left(\frac{1}{n}\right) + 2 \cdot \left(\frac{1}{n}\right)^2 + 3 \cdot \left(\frac{1}{n}\right)^3 + \cdots\right] \\
&= 1 + \frac{1}{n} + \left(\frac{1}{n}\right)^2 + \cdots \\
&= \frac{1}{1 - 1/n} \\
&= \frac{n}{n-1}
\end{aligned}$$

Now consider X_3, the number of cereal boxes to buy to get the third coupon after you have the first two coupons.

$$\begin{aligned}
\mathrm{E}[X_3] &= 1 \cdot \frac{n-2}{n} + 2 \cdot \left(\frac{2}{n}\right) \frac{n-2}{n} + 3 \cdot \left(\frac{2}{n}\right)^2 \frac{n-2}{n} + \cdots \\
&= \frac{1}{1 - 2/n} \\
&= \frac{n}{n-2}
\end{aligned}$$

In general, consider X_i, the number of cereal boxes to buy get the i type of coupon after you have collected the first $i-1$ distinct type of coupons. I get

$$\begin{aligned}
\mathrm{E}[X_i] &= 1 \cdot \frac{n-i+1}{n} + 2 \cdot \left(\frac{i-1}{n}\right) \frac{n-i+1}{n} + 3 \cdot \left(\frac{i-1}{n}\right)^2 \frac{n-i+1}{n} + \cdots \\
&= \frac{1}{1 - (i-1)/n} \\
&= \frac{n}{n-i+1}
\end{aligned}$$

Let $X = X_1 + X_2 + X_3 + \cdots + X_n$, i.e., the number of boxes to buy to get all the n distinct types of coupon. Then

$$\mathrm{E}[X] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \sum_{i=1}^{n} \frac{1}{n-i+1}$$

Note that the sum is

$$\begin{aligned}\sum_{i=1}^{n} \frac{1}{n-i+1} &= \frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots + \frac{1}{1} \\ &= 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \\ &= \sum_{i=1}^{n} \frac{1}{i}\end{aligned}$$

Hence

$$\mathrm{E}[X] = n \sum_{i=1}^{n} \frac{1}{i}$$

□

Note that for this example, all the $\mathrm{E}[X_i]$'s are different.

Aside. The sum

$$\sum_{i=1}^{n} \frac{1}{i}$$

is called the n–the **Harmonic number** and is denoted by H_n. It's known that H_n is very close to $\ln n$ ($\ln$ is log to base e). Specifically

Harmonic number

$$\lim_{n\to\infty} (H_n - \ln n) = 0.5772\cdots$$

is known to be a constant. This constant

$$\gamma = \lim_{n\to\infty} (H_n - \ln n)$$

is called the **Euler-Mascheroni constant**. γ appears in many areas of math, CS, and even physics. Very little is known about γ. It is not even known if γ is irrational – this is a very difficult open problem. It is known that if γ is rational, then the denominator of γ is greater than 10^{242080}.

Euler-Mascheroni constant

Exercise 9.10.12. How many rolls of a fair die do you need to get all the possible faces? Write a program to check your answer. What if you have an n–sided die? □

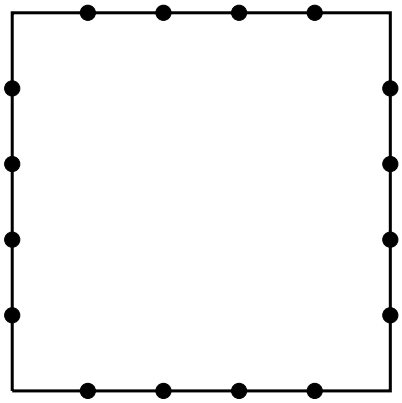
Exercise 9.10.13. Suppose again you want to roll a die until you get all the possible faces. Now note that to get your first number, you need only one roll. Suppose you get a Four. For your second roll, you might get a different

number with one roll, but you might not. But suppose (because I'm nice), I let you have two tries to get a different number. Suppose on your first roll to get a number different from Four, you get Two in just one roll – lucky you. For your third number (different from Four and Two), I give you three tries. Suppose you rolled three times and you got Four, Two, Two – you can't get a different number in three tries. Too bad. In that case you have to start all over again, i.e., your Four and Two does not count anymore. How many rolls do you need to get all the six different faces? What if you have an n–sided fair die? □

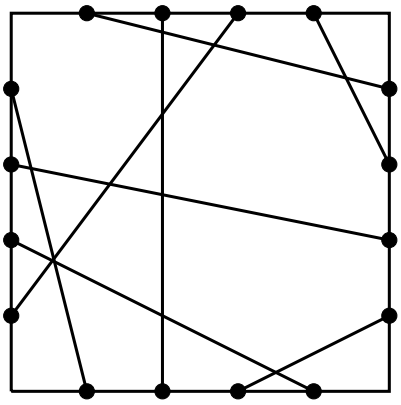
Exercise 9.10.14. If you toss a coin several times to get HTHHHHTHTTHT you get the following runs: H, T, HHHH, T, H, TT, H, T. (I'm writing H for Head and T for Tail.) What is the average number of runs if a fair coin is tossed n times?

Exercise 9.10.15. If you stack (randomly) n dice on top of each other, what is the probability that the top face of the dice form a palindrome? What is the number of palindromes? □

Exercise 9.10.16. Draw a square like this with n points on each side:



(This picture has $n = 4$.) Pairs of points are randomly chosen to be connected by a line segment. Each point is on exactly one line segment. A point can only be paired with a point on a different side of the square. Here's an example:



- How many points of intersection are there?
- How many polygons are formed?

Exercise 9.10.17. What is the average number of "shuffles" you need in order to "fully randomize" a desk of cards? Make sure you define your "shuffle" operation and what you mean by a "random" deck of cards. For instance a simple shuffle is to cut the deck: Assuming the deck is $[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$, you pick a random number such as 6, you cut the deck to get $[0, 1, 2, 3, 4, 5]$ and $[6, 7, 8, 9]$ and then get $[6, 7, 8, 9, 0, 1, 2, 3, 4, 5]$.

Exercise 9.10.18. You are at a firing range:

```
x = [-,-,-,-,-,-,-,-,-,-,-,-,-,-,-,-,-,-,-,-]
```

which is made up of glass. You are not good at shooting and so you fired k shots randomly:

```
x = [-,X,-,X,-,-,-,X,X,-,-,X,-,-,X,-]
```

The glass then cracked from each shot to the nearest other shot:

```
x = [-,X,X,X,-,-,-,X,X,-,-,X,X,X,X,-]
```

What fraction of the glass is cracked? (Fraction = ratio of `X`s to the size of array)?

File: discrete-probability/projection.tex

9.11 Projection

Somewhat similar is the case where you use random variables as projection onto a coordinate for the case when the sample space is a product of space spaces. For instance consider the random experiment of tossing a coin and rolling a die. The sample space is

$$S = \{\text{Head}, \text{Tail}\} \times \{\text{One}, \text{Two}, ..., \text{Six}\}$$

We can define the random variable X to be

$$X : S \to \{\text{Head}, \text{Tail}\}$$

where

$$X((x, y)) = x$$

and Y to be

$$Y : S \to \{\text{One}, ..., \text{Six}\}$$

where

$$Y((x, y)) = y$$

Or even better, instead of X, I use Toss and instead of Y, I use Roll.

(Aside. Note that the outcome looks like (x, y). So the correct way to write the X of (x, y) is $X((x, y))$. But it's also common to write it as $X(x, y)$.)

This is like labelings.

To make things more readable, let me define the random variable Toss to be the same as X and the random variable Roll to be the same as the random variable Y. The random variable Toss labels the outcome (Head, Two) as Head – i.e., it simply labels (x, y) as x.

Now consider probabilities. It should be clear that, assuming the coin and die are both fair that the pdf is uniform:

$$p((x, y)) = 1/12$$

The probability

$$\Pr[\text{Toss} = \text{Head}]$$

is simply

$$p(\{(x, y) \in S \mid \text{Toss}(x, y) = \text{Head}\})$$

which is

$$p(\{(\text{Head}, y) \mid y \in \{\text{One}, ..., \text{Six}\}\})$$

This is the probability of tossing a coin and rolling a die and the coin toss happens to be head.

Exercise 9.11.1.

1. What is Pr[Toss = Head]?
2. What is Pr[Roll = One]?
3. Suppose $p_{\text{Toss}} : \{\text{Head}, \text{Tail}\} \to \mathbb{R}$ is the pdf of the random experiment of the toss of a fair coin. What can you tell me about the relationship between function p_{Toss} and the values Pr[Toss = Head] and Pr[Toss = Tail]? Make sure you see that p_{Toss} and Pr[Toss = •] are functions on different sample spaces.
4. Suppose $p_{\text{Roll}} : \{\text{One}, ..., \text{Six}\} \to \mathbb{R}$ is the pdf of the random experiment of rolling a fair die. What can you tell me about the relationship between function p_{Roll} and the values Pr[Roll = One], $\cdots$, Pr[Roll = Six]? Make sure you see that p_{Roll} and Pr[Roll = •] are functions on different sample spaces.

The point of the above is that

$$p_{\text{Toss}}(x) = \Pr[\text{Toss} = x]$$

The left-hand side is about the sample space $\{\text{Head}, \text{Tail}\}$ whereas on the right-hand side, the sample space is $\{\text{Head}, \text{Tail}\} \times \{\text{One}, \ldots, \text{Six}\}$.

File: discrete-probability/bernoulli-trial.tex

9.12 Bernoulli trials and Binomial distribution

At this point, we have the basic concept of the probability attached to a random experiment. I have also talked about an experiment that is broken up into two independent random experiment – this is when the pdf is a product of two pdfs.

Now I want to talk about the case where a random experiment involves performing a *sequence* of the *same* random experiment. The sequence need not be made up of two experiments. Usually there's no limit on the size of the sequence. In fact, I am usually interested in questions like "How many times do I need to execute this random experiment until a goal is reached?"

To be specific, consider the following question:

"If there's a 32% change of making \$1 when I buy one google stock on Monday at 10:42AM and sell it on the same day at 2:45PM, what is the chance of me making \$100 if I buy and sell one google stock according to the above day and times for 200 consecutive Mondays when the stock market is open". Or: "How many consecutive Mondays of trading do I need to execute before I made 20 days of gains?"

A **Bernoulli trial** is a random experiment with two outcomes. This term is used especially when the Bernoulli trials are performed in a sequence such that the trials are mutually independent.

For instance suppose I have a biased coin: the probability of getting head is 1/3. What then is the probability of getting 4 heads when the coin is tossed 10 times?

Let me put everything into proper mathematical notation. Let pdf $p_i : S_i \to [0, 1]$ denote the i–th time you are tossing the coin. Note that

$$S_i = \{\textsc{Head}, \textsc{Tail}\}$$

and

$$p_i(\textsc{Head}) = 1/3, \quad p_i(\textsc{Tail}) = 2/3$$

(they are all the same, i.s., $S_i = S_j$ and $p_i = p_j$). The experiments are mutually independent. Therefore if $p : S \to [0, 1]$ is the pdf for the experiment of tossing

the coin 10 times where $S = S_1 \times \cdots \times S_{10}$, then

$$p(x_1, \ldots, x_{10}) = p_1(x_1) \cdots p_{10}(x_{10})$$

(Note that the probability is a product since the i–th toss is independent of the j–th toss for $i \neq j$.) I am interested in the event

$$A = \{(x_1, \ldots, x_{10}) \in S \mid \text{exactly four of the } x_i\text{'s are } \textsc{Head}\}$$

Note that

$$|A| = \binom{10}{4}$$

and each element of A has the same probability as the case where the first four tosses are heads:

$$\begin{aligned} &p(\textsc{Head}, \textsc{Head}, \textsc{Head}, \textsc{Head}, \textsc{Tail}, \ldots, \textsc{Tail}) \qquad (4 \text{ heads, } 6 \text{ tails}) \\ &= p_1(\textsc{Head}) p_2(\textsc{Head}) p_3(\textsc{Head}) p_4(\textsc{Tail}) \cdots p_{10}(\textsc{Tail}) \\ &= (1/3)^4 (2/3)^4 \end{aligned}$$

Therefore the required probability is

$$p(A) = \binom{10}{4} (1/3)^4 (2/3)^6$$

Because a Bernoulli trial has two outcomes, it's common to call one of the outcomes a **success** and the other a **failure**.

success
failure

In general, you see right away that:

Theorem 9.12.1. *If the probability of the success of a Bernoulli trial is p, then the probability of having k successes when performing n of the Bernoulli trials is given by*

$$\binom{n}{k} p^k (1-p)^{n-k}$$

□

Frequently you will see the following notations:

$$B_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

or

$$B(n, p; k) = \binom{n}{k} p^k (1-p)^{n-k}$$

or

$$B(n,p,k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Frequently the n and p are fixed and k is considered the variable of the pdf $B(n,p)$.

Formally, define the pdf of a **Bernoulli trial** as Bernoulli trial

$$p_{\text{Bernoulli}} : \{\text{Success}, \text{Failure}\} \to [0,1]$$

Instead of using the outcomes Success and Failure, it's useful to use 1 and 0. In other words, it's useful to have a Bernoulli trial random variable defined as

$$X_{\text{Bernoulli}} : \{\text{Success}, \text{Failure}\} \to \{1,0\}$$

where

$$X_{\text{Bernoulli}}(\text{Success}) = 1, \quad X_{\text{Bernoulli}}(\text{Failure}) = 0$$

especially since for the corresponding Binomial distribution I'll need to count the number of successes. In other words $X_{\text{Bernoulli}}$ is an indicator random variable of Success. All the above notations are for a single Bernoulli trial. Now we define the **Binomial distribution**. Binomial distribution

Associated with a given Bernoulli trial $p_{\text{Bernoulli}}$, we define the n–fold product pdf, the **Binomial distribution** Binomial distribution

$$p_{B(n,p)} : \{\text{Success}, \text{Failure}\}^n \to [0,1]$$

as

$$p_{B(n,p)}(x_0, x_1, \ldots, x_{n-1}) = p_0(x_0) \cdots p_{n-1}(x_{n-1})$$

where p_i is the pdf of the i–th Bernoulli trial where the "p" is the probability of success of the Bernoulli trail. Of course $p_i = p_{\text{Bernoulli}}$, i.e.,

$$p_{B(n,p)}(x_0, x_1, \ldots, x_{n-1}) = p_{\text{Bernoulli}}(x_0) \cdots p_{\text{Bernoulli}}(x_{n-1})$$

Let X_i be the indicator random variable for a success for the i–th Bernoulli trial, i.e.,

$$X_i(x_0, ..., x_{n-1}) = \begin{cases} 1 & \text{if } x_i = \text{Success} \\ 0 & \text{otherwise} \end{cases}$$

In other words

$$X_i(x_0, ..., x_{n-1}) = X_{\text{Bernoulli}}(x_i)$$

Next define the random variable

$$X_{B(n,p)} = \sum_{i=0}^{n-1} X_i$$

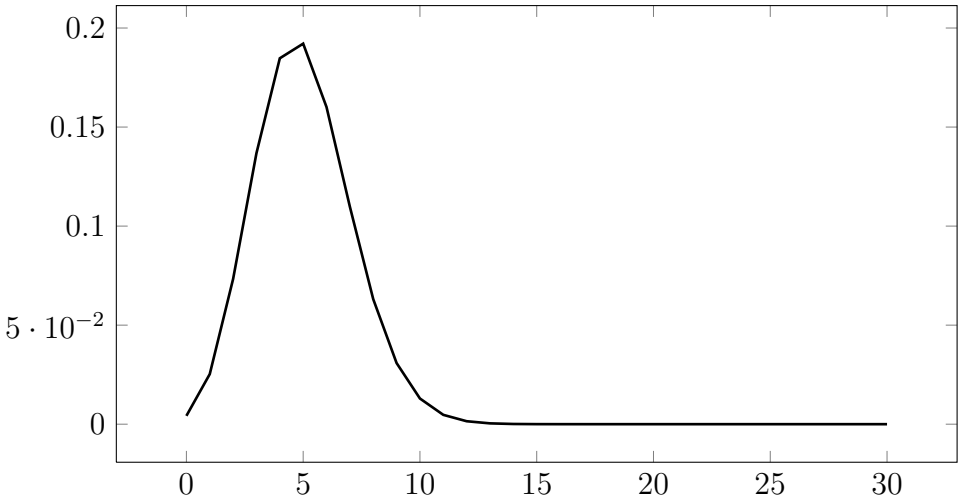
Then

$$B_{n,p}(k) = \Pr[X_{B(n,p)} = k]$$

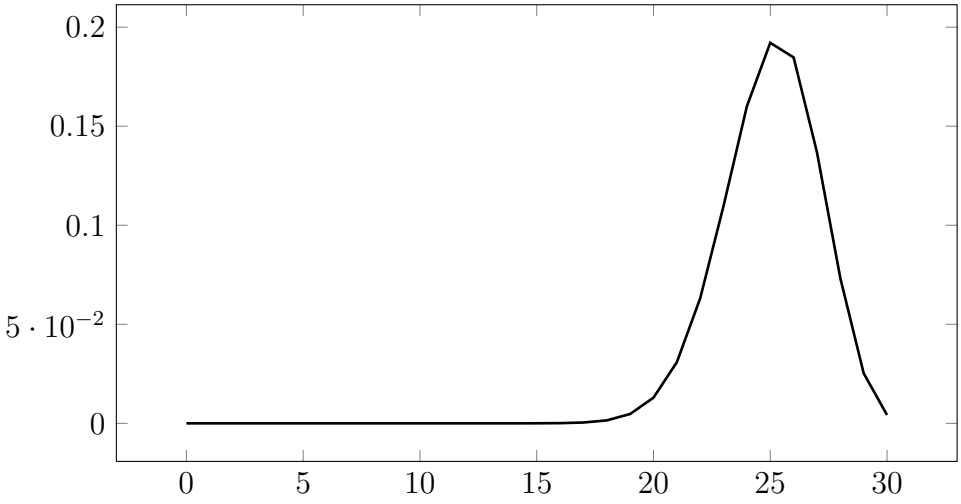
With these notation, the above theorem says

$$B_{n,p}(k) = \Pr[X_{B(n,p)} = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

Here's a plot of $B(30, 0.1)$:



and here's a plot when $p = 5/6$:



Exercise 9.12.1.

1. If a biased coin which has a probability of getting a head is 1/4 is tossed 8 times, what is the probability of getting exactly 3 tails?
2. What if I changed the above to "3 or less tails"?
3. What if I changed the above to "exactly 3 consecutive tails"?

□

Exercise 9.12.2. A sequence of 10 bits are generated so that the probability of producing a 0 is 0.1.

1. What is the probability of getting a bit string that begins with three 0s. Assume that the generation of the bits are mutually independent.
2. What if I change the above to "a bit string with three 0s"?
3. What if I change the above to "a bit string with at least three 0s"?
4. What if I change the above to "a bit string with at least three 0s" and the sequence is not a sequence of length 10, but of length 1000?
5. What if I change the above to "a bit string with no three consecutive 0s" and the sequence is not a sequence of length 10, but of length 1000?

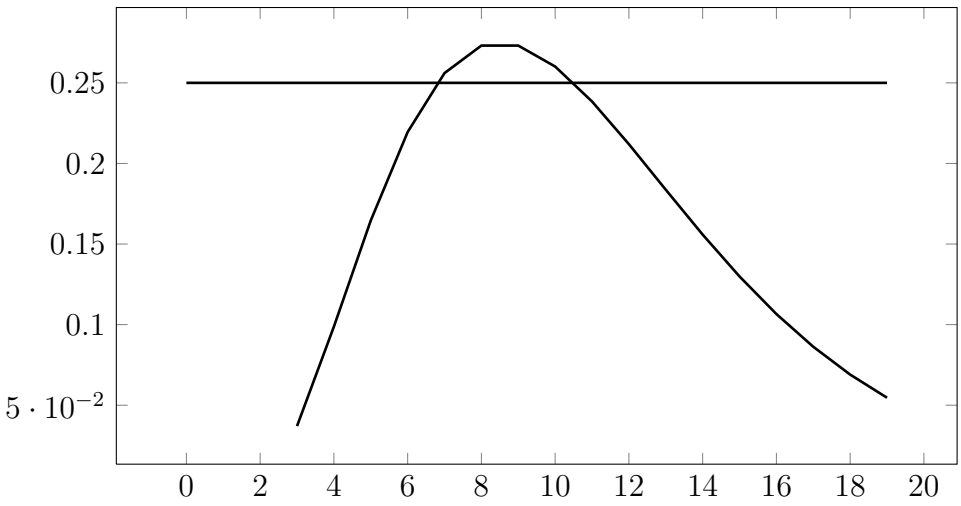
□

Exercise 9.12.3. A sequence of 10 bits are generated so that the probability of producing a 0 is 0.1. What is the probability of getting a bit string that begins with three 0s and has five 0s in total. Assume that the generation of the bits are mutually independent. □

Exercise 9.12.4. How many coin tosses do you need so that the probability of getting 3 heads is at least 0.25. Assume the probability of getting a head is 1/3. □

SOLUTION.
Basically we want to find n such that

$$\binom{n}{3}(1/3)^3(2/3)^{n-3} \geq 0.25$$

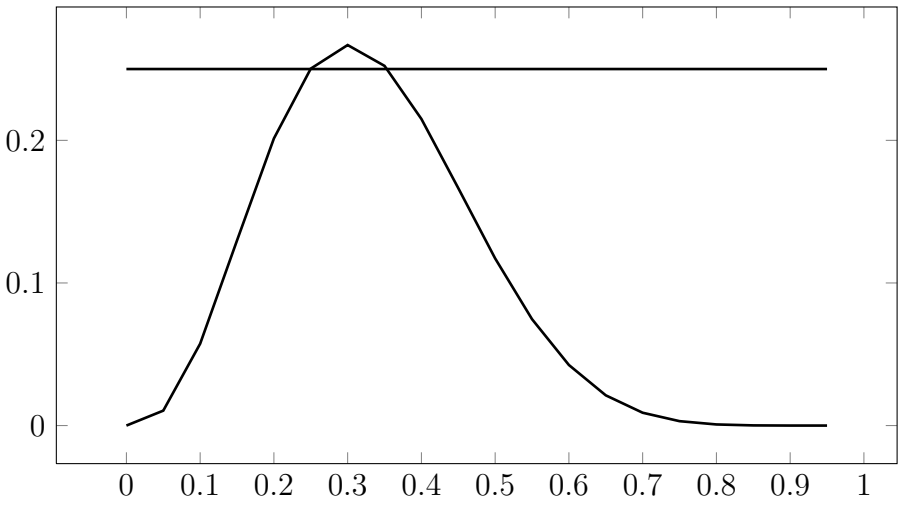


From the graph, $n = 7, 8, 9, 10$. □

Exercise 9.12.5. If the probability of getting 3 heads with 10 coin tosses is > 0.5, what is range of the probability of getting a head for this coin? □

SOLUTION.
Basically we want to find n such that

$$\binom{10}{3} p^3 (1-p)^{n-3} \geq 0.5$$



From the graph, $n = 7, 8, 9, 10$. □

Exercise 9.12.6. You have decided to set up a game table as a fund raising event for the CS club. To play the game, a player has to pay you \$1. After paying you \$1, he/she will roll a die $n = 7$ times. If he/she gets at least 2 sixes, he/she will get \$100. Of course you want to make money (duh) – what kind of fund raising event wants to lose money?? So you have decided to lower the chance of the player getting at least $k = 2$ sixes in $n = 7$ rolls. You found a 3D printer and a program that will allow you to build a loaded die. What is the lowest probability to use if you want to break even? (Of course you do want to raise some money. If you make it impossible to get a six, then after just a couple of games, people would know that the die is rigged. So you want to know your break even point and then slightly lower the probability of getting a six.)

The probability of getting at least 2 sixes is

$$B(n,p;2) + B(n,p;3) + B(n,p;4) + \cdots + B(n,p;7)$$

Therefore the student will earn

$$(B(n,p;2) + B(n,p;3) + B(n,p;4) + \cdots + B(n,p;7)) \times 100 - 1$$

(The "-1" is because the student has to pay \$1 to play.) Of course to make money, you will want

$$(B(n,p;2) + B(n,p;3) + B(n,p;4) + \cdots + B(n,p;7)) \times 100 - 1 < 0$$

i.e.,

$$B(n,p;2) + B(n,p;3) + B(n,p;4) + \cdots + B(n,p;7) < 1/100$$

It's easier to use this:

$$1 - B(n,p;0) - B(n,p;1) - B(n,p;2) < 1/100$$

and therefore we want

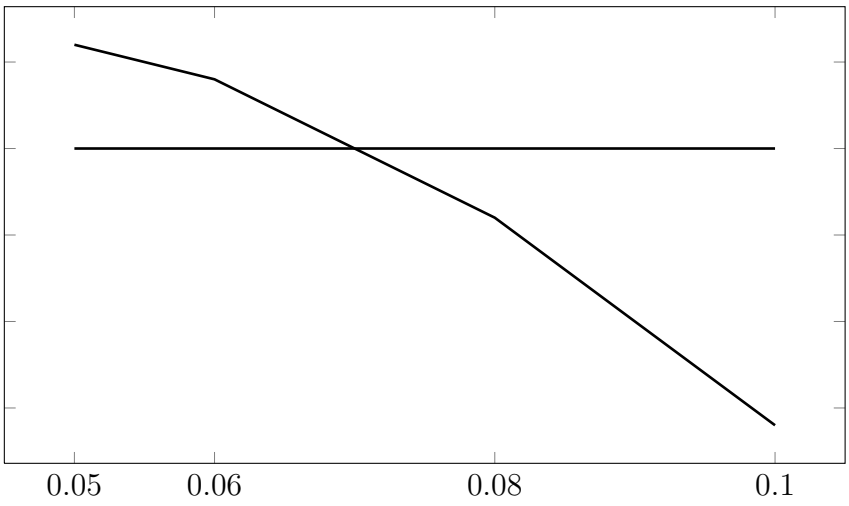
$$99/100 < B(n,p;0) + B(n,p;1) + B(n,p;2)$$

which is the same as

$$99/100 < \binom{7}{0}p^0(1-p)^{7-0} + \binom{7}{1}p^1(1-p)^{7-1} + \binom{7}{2}p^2(1-p)^{7-2}$$

The following is a graph for different values of p:

Zooming in on the part of the graph where x is in $[0.05, 0.1]$, we get



Therefore the break even value for p is approximately 0.07. (For a fair die, the probability for getting a six is $1/6 = 0.167$.)

We now compute our gains for a few values of p near 0.07. If $p = 0.06$, then the player would make

$$(1 - B(7, 0.06; 0) - B(7, 0.06; 1) - B(7, 0.06; 2)) \times 100 - 1 = -0.3706...$$

which means that we gain \$0.37 per game. If $p = 0.07$, then the player would make

$$(1 - B(7, 0.07; 0) - B(7, 0.07; 1) - B(7, 0.07; 2)) \times 100 - 1 = -0.0312...$$

which means that we gain \$0.03 per game. If $p = 0.08$, then the player would make

$$(1 - B(7, 0.08; 0) - B(7, 0.08; 1) - B(7, 0.08; 2)) \times 100 - 1 = 0.4014....$$

which means that we would lose \$0.40 per game.

So, for instance, if we use $p = 0.07$ and a total of 100 games were played, we will collect \$37.00.

[TO TIDY UP]

[PUT THIS PROBLEM SOMEWHERE]

Here are problems of the form "How many times (on the average, of course) must you perform an experiment in order to get ...".

Exercise 9.12.7. What is the expected number of flips of a coin that shows a head with probability p if we want to get k heads?

```
import random; random.seed()

def get_flips(n, k):
    """
    Probability of getting a head is 1.0 / n
    Returns the number of flips of coin to reach k heads.
    """
    flips = 0
    heads = 0
    while heads < k:
        flips += 1
        face = random.randrange(n)
        if face == 0:
            heads += 1
    return flips

n = input("n where p(HEAD) is 1/n: ")
k = ("number of heads to reach: ")

N = 1000 # number of experiments
total_flips = 0
for i in range(N):
    flips = get_flips(n, k)
    total_flips += flips
    print "flips:", str(flips).rjust(4), \
          " E:", float(total_flips) / (i + 1)
```

I ran the above for $p = 1/10$ and $k = 10$ and got about 100. When I ran it at $p = /15$ and $k = 20$, I got approximately 300. So, experimentally, it seems that the expected number is

$$\frac{k}{p}$$

Here's the solution: If you toss the coin once, the expected number of heads is p. If you toss the coin twice, the expected number of heads is $p + p = 2p$. In general if X_i is the indicator random variable for the i–th toss being a head, then we want to find i such that

$$E[X_1 + X_2 + \cdots + X_i] = k$$

Since

$$E[X_1 + X_2 + \cdots + X_i] = E[X_1] + E[X_2] + \cdots + E[X_i] = ip$$

So to get k heads,

$$ip = k$$

and therefore

$$i = k/p$$

i.e., on the average, the number of tosses to get k heads is k/p. □

Exercise 9.12.8. On the average, how many times must you a roll a die to get k sixes? □

Exercise 9.12.9. On the average, how many times must you a roll a die to get 3 sixes? k consecutive sixes? □

Exercise 9.12.10. On the average, how many times must you toss a coin to not get alternating HTHT... or THTH...? □

Exercise 9.12.11. You write a program to generate 10 random numbers chosen from $0, 1, 2, ..., 99$. How many times do you need to run the program (on the average) to get an ascending sequence? □

Exercise 9.12.12. You have a random sequence of 10 distinct numbers. You randomly pick two numbers and swap them if they are no in ascending order. On the average, how many times must you perform the above swapping operation to sort the numbers in ascending order? □

Exercise 9.12.13. You have a regular n–gon. You randomly pick two pairs of vertices of the n–gon and join them. On the average, how many pairs do you need to join in order to get at least a point of intersection not on a vertex?

File: discrete-probability/bayes/bayes.tex

9.13 Baye's Theorem

Recall that earlier, I talked about a random experiment that is made up of performing two independent experiments. Now suppose I have two random experiments which are not independent. For instance, I have two boxes with the following contents:

- box 1: 1 green balls and 3 red balls
- box 2: 2 green balls and 1 red balls

The random experiment involve choosing one box and then picking a ball from the box. Notice that the event of picking a red ball (the second random experiment) depends on the first (the first random experiment).

Recall the we define

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

This is the probability of A given B has occured. Note that we also have

$$p(B \mid A) = \frac{p(B \cap A)}{p(A)}$$

Therefore we get

$$\begin{aligned} p(A \mid B) \cdot p(B) &= p(A \cap B) \\ p(B \mid A) \cdot p(A) &= p(B \cap A) \end{aligned}$$

Since $A \cap B = B \cap A$, we get the following immediately:

$$p(B \mid A) \cdot p(A) = p(A \mid B) \cdot p(B)$$

Notice the beautiful symmetry?

Beauty aside ... what's the whole point of the above? Well, sometimes you will find that given A and B, it's pretty easy to compute $p(A \mid B)$ instead of $p(B \mid A)$.

Let me give you an example ...

Example 9.13.1. I have two boxes with the following contents:

- box 1: 1 green balls and 3 red balls
- box 2: 2 green balls and 1 red balls

The random experiment involve choosing one box and then picking a ball from the box. The possible outcomes are:

- $(1, G)$
- $(1, R)$
- $(2, G)$
- $(2, R)$

where the first coordinate represents the box chosen and the second is the color of the ball chosen from the box indicated by the first coordinate. Now let me define two events:

$$B = \text{I choose box 1 (and after that I pick a ball from box 1)}$$
$$A = \text{I pick a red ball (after I picked a box)}$$

Now let's think about $p(B \mid A)$ and $p(A \mid B)$.

$p(A \mid B)$ is easy: this is the probability of choosing a red ball if I picked box 1. Why is this easy? Well, if I picked box 1, then the probability of picking a red ball must be $3/4$ since there are 3 red balls out of altogether 4 balls in box 1.

$$p(A \mid B) = \frac{3}{4}$$

Easy! Formally $A = \{(1, R), (2, R)\}$ and $B = \{(1, R), (1, G)\}$. Therefore

$$p(A \mid B) = \frac{p(\{(1, R)\})}{p(\{(1, R), (1, G)\})} = \frac{1/2 \cdot 3/4}{1/2 \cdot 3/4 + 1/2 \cdot 1/4} = \frac{3}{4}$$

[Include section on experiment which is a sequence of experiments which are dependent.

$$p(B \mid A) = \frac{p(\{(1, R)\})}{p(\{(1, R), (2, R)\})} = \frac{1/2 \cdot 3/4}{1/2 \cdot 3/4 + 1/2 \cdot 1/3} = \frac{3/4}{3/4 + 1/3}$$

]

However it's not immediately clear how to compute $p(B \mid A)$. This is the probability of choosing box 1 if I picked a red ball. The earlier result:

$$p(B \mid A) \cdot p(A) = p(A \mid B) \cdot p(B)$$

tells us that there's a connnection between $p(A \mid B)$ and $p(B \mid A)$. Of course I just said that $p(A \mid B)$ is 3/4. Furthermore $p(B)$ is 1/2. The above becomes:

$$p(B \mid A)p(A) = \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8}$$

I'm done if I can figure out $p(A)$, since from the above I have

$$p(B \mid A) = \frac{1}{p(A)} \cdot \frac{3}{8}$$

$p(A)$ is the probability of choosing a red ball, This is rather complicated since there are two separate scenarios: a red ball from box 1 or a red ball from box 2. Now what?

Well ... I just have to consider the two separate scenarios! Duh!

I have to consider the scenario where the red ball is taken from box 1 and then the scenario where the red ball is taken from box 2. Note that the two scenario are disjoint. Mathematically, I'm simply computing $p(A)$ by doing this (informally speaking):

$$p(A) = p((\text{box 1 case}) \cap A) + p((\text{box 2 case}) \cap A)$$

Using B, this is just

$$p(A) = p(B \cap A) + p(\overline{B} \cap A)$$

Note that the two events $B \cap A$ and $\overline{B} \cap A$ are disjoint and the union is indeed A. In other words I'm using the general fact that if events E_1 and E_2 are disjoint, then

$$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2) = p(E_1) + p(E_2) - 0 = p(E_1) + p(E_2)$$

Now let's compute $p(A \cap B)$. Note this is the probability that I picked box 1 *and* I picked a red ball. This is *not* the probability that I picked a red ball *given* that I picked box 1.

$$p(A \cap B) = p(A \mid B)p(B) = \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8}$$

The other quantity is

$$p(A \cap \overline{B}) = p(A \mid \overline{B})p(\overline{B}) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

Putting things together we get

$$p(A) = p(B \cap A) + p(\overline{B} \cap A) = \frac{3}{8} + \frac{1}{6} = \frac{13}{24}$$

Putting everything together we get

$$p(B \mid A) = \frac{1}{p(A)} \cdot \frac{3}{8} = \frac{24}{13} \cdot \frac{3}{8} = \frac{9}{13}$$

Let me finish by saying that sometimes if you want to compute $p(B \mid A)$, it might be easier to use this

$$p(B \mid A) \cdot p(A) = p(A \mid B) \cdot p(B)$$

assuming that it's easier to compute $p(A \mid B)$. And for the computation of $p(A)$, in some cases, it might be easier to use

$$p(A) = p(A \cap B) + p(A \cap \overline{B})$$

where the quantities on the right can be computed using

$$p(A \cap B) = p(A \mid B) \cdot p(B)$$
$$p(A \cap \overline{B}) = p(A \mid \overline{B}) \cdot p(\overline{B})$$

If you put everything into the expression for $p(B \mid A)$, you get this

$$p(B \mid A) = \frac{1}{p(A \mid B)p(B) + p(A \mid \overline{B})p(\overline{B})} \cdot p(A \mid B) \cdot p(B)$$

Before I state the computational technique as a theorem (the Baye's Theorem) let me compute the same probability the naive way, staight from definition by listing the outcomes.

In that case the sample space of our experiment is:

```
1, G         B
1, R    A    B
1, R    A    B
1, R    A    B
2, G
2, G
2, R    A
```

(The first column is the box, the second is the ball. The A and B indicates which event the row belongs to.) However, note that the number of items is not correct. For instance note that probability of choosing box 1 must be 0.5. So let's be duplicate the first 4 lines 3 times and the last 3 line 4 times:

```
1, G         B
1, R    A    B
1, R    A    B
1, R    A    B
1, G         B
1, R    A    B
1, R    A    B
1, R    A    B
1, G         B
1, R    A    B
1, R    A    B
1, R    A    B
2, G
2, G
2, R    A
2, G
2, G
2, R    A
2, G
2, G
2, R    A
2, G
2, G
2, R    A
```

The probability we want is $p(B \mid A)$ which is defined to be $p(B \cap A)/p(A)$. Now, from the above table, $p(B \cap A) = 9/24 = 3/8$. Also, $p(A) = 13/24$. Therefore $p(B \mid A) = (3/8)/(13/24) = 9/13$.

Our brute force computation does match our computation above.

Now I'm going to give you a *third* method using a simulation. This is actually a random experiment to get an approximation of the probability that we want. Because it's a simulation, it won't be exact. However the more random trials we simulate the closer we get to the actual probability. The code below explains itself. I'm going to simulate 1000000 random trials of the given experiment.

```
import random;random.seed()
count1 = 0 # count of outcomes in A and B
```

```
count2 = 0 # count of outcomes in A

N = 10000000

for i in range(N):
    # randomly pick box 1 or box 2
    box = random.randrange(1, 3)
    # randomly pick ball as 'G' or 'R'
    # based on which box was chosen
    if box == 1:
        ball = random.choice(['G', 'R', 'R', 'R'])
    else:
        ball = random.choice(['G', 'G', 'R'])
    if ball == 'R': # outcome is in A
        count2 += 1
        if box == 1: # outcome is in B
            count1 += 1

    #if i % 100000 == 0 and count2 != 0:
print (float(count1) / count2)
```

The output is

```
0.692257273456
```

and

$$\frac{9}{13} = 0.6923076923076923...$$

so the simulation agrees with out exact computation by up to 3 decimal places.

Exercise 9.13.1. I toss a fair coin. If the outcome is head, I toss an unfair coin with a probability of getting a head being 1/3. If the outcome of the first coin is tail, I toss another unfair coin with a probability of getting a head being 1/5. What is the probability that the first outcome is a head if the second outcome is a tail?

Let A be the event that the first outcome is a head and B be the event that the second outcome is a tail. We are given:

$$\begin{aligned} p(A) = p(\overline{A}) &= \frac{1}{2} \\ p(B \mid A) &= \frac{2}{3} \\ p(B \mid \overline{A}) &= \frac{4}{5} \end{aligned}$$

We want $p(A \mid B)$. We have

$$p(A \mid B) \cdot p(B) = p(B \mid A) \cdot p(A)$$

which implies

$$\begin{aligned} p(A \mid B) &= \frac{1}{p(B)} \cdot p(B \mid A) \cdot p(A) \\ &= \frac{1}{p(B)} \cdot \frac{2}{3} \cdot \frac{1}{2} \\ &= \frac{1}{p(B)} \cdot \frac{1}{3} \end{aligned}$$

Now

$$\begin{aligned} p(B) &= p(B \cap A) + p(B \cap \overline{A}) \\ &= p(B \mid A) \cdot p(A) + p(B \mid \overline{A}) \cdot p(\overline{A}) \\ &= \frac{2}{3} \cdot \frac{1}{2} + \frac{4}{5} \cdot \frac{1}{2} \\ &= \frac{1}{3} + \frac{2}{5} \\ &= \frac{11}{15} \end{aligned}$$

Therefore

$$p(A \mid B) = \frac{1}{p(B)} \cdot \frac{1}{3} = \frac{15}{11} \cdot \frac{1}{3} = \frac{5}{11}$$

□

Exercise 9.13.2. I have three boxes.

1. Box 1: 3 red balls, 2 white balls
2. Box 2: 4 red balls, 2 white balls
3. Box 3: 2 red balls, 3 white balls

This is what I'm going to do. I first draw a ball from Box 1. If the ball from Box 1 is red, I pick a ball from Box 2. If the ball from Box 1 is white, I pick a ball from Box 3.

For this problem, I will deliberately use random variable notation instead of events notation. I will leave it to you to define the random variable formally. Just from the name of the random variable, you should see quickly how to define the random variable formally and correctly.

Here are the questions:

1. What is $\Pr[\textsc{Ball2} = \textsc{Red} \mid \textsc{Ball1} = \textsc{Red}]$?
2. What is $\Pr[\textsc{Ball1} = \textsc{Red} \mid \textsc{Ball2} = \textsc{Red}]$?

The first probability is easy:

$$\Pr[\textsc{Ball2} = \textsc{Red} \mid \textsc{Ball1} = \textsc{Red}] = \frac{4}{6} = \frac{2}{3}$$

For the second probability we use Baye's theorem:

$$\begin{aligned}&\Pr[\textsc{Ball1} = \textsc{Red} \mid \textsc{Ball2} = \textsc{Red}] \cdot \Pr[\textsc{Ball2} = \textsc{Red}] \\ &\quad = \Pr[\textsc{Ball2} = \textsc{Red} \mid \textsc{Ball1} = \textsc{Red}] \cdot \Pr[\textsc{Ball1} = \textsc{Red}]\end{aligned}$$

which gives us the following immediately:

$$\Pr[\textsc{Ball1} = \textsc{Red} \mid \textsc{Ball2} = \textsc{Red}] \cdot \Pr[\textsc{Ball2} = \textsc{Red}] = \frac{2}{3} \cdot \frac{3}{5} = \frac{2}{5}$$

and therefore

$$\Pr[\textsc{Ball1} = \textsc{Red} \mid \textsc{Ball2} = \textsc{Red}] = \frac{1}{\Pr[\textsc{Ball2} = \textsc{Red}]} \cdot \frac{2}{5}$$

Now

$$\begin{aligned}
&\Pr[\textsc{Ball2} = \textsc{Red}] \\
&\quad = \Pr[\textsc{Ball2} = \textsc{Red}, \textsc{Ball1} = \textsc{Red}] + \Pr[\textsc{Ball2} = \textsc{Red}, \textsc{Ball1} \neq \textsc{Red}] \\
&\quad = \Pr[\textsc{Ball2} = \textsc{Red} \mid \textsc{Ball1} = \textsc{Red}] \cdot \Pr[\textsc{Ball1} = \textsc{Red}] \\
&\qquad + \Pr[\textsc{Ball2} = \textsc{Red} \mid \textsc{Ball1} \neq \textsc{Red}] \cdot \Pr[\textsc{Ball1} \neq \textsc{Red}] \\
&\quad = \frac{4}{6} \cdot \frac{3}{5} + \frac{2}{5} \cdot \frac{2}{5} \\
&\quad = \frac{2}{5} + \frac{4}{25} \\
&\quad = \frac{14}{25}
\end{aligned}$$

Therefore

$$\begin{aligned}
\Pr[\textsc{Ball1} = \textsc{Red} \mid \textsc{Ball2} = \textsc{Red}] &= \frac{1}{\Pr[\textsc{Ball2} = \textsc{Red}]} \cdot \frac{2}{5} \\
&= \frac{1}{14/25} \cdot \frac{2}{5} \\
&= \frac{25}{14} \cdot \frac{2}{5} \\
&= \frac{5}{7}
\end{aligned}$$

Here's a simulation:

```
import random;random.seed()
count1 = 0 # count of ball 1 = 'R'
count2 = 0 # count of ball 2 = 'R'

for i in xrange(10000000):

    ball1 = random.choice(['R','R','R','W','W'])

    if ball1 == 'R' :
        ball2 = random.choice(['R','R','R','R','W','W'])
    else:
        ball2 = random.choice(['R','R','W','W','W'])

    if ball2 == 'R':
        count2 += 1
        if ball1 == 'R':
            count1 += 1

print float(count1) / count2
```

The output is

```
0.714076007643
```

which agrees with our computation up to 3 decimal places:

$$\frac{5}{7} = 0.7142857142857143...$$

□

Note that the first question is easy: Given that Ball1 is red, the ball we draw would be from Box 2 and the probability of drawing a red is $\frac{4}{6} = \frac{2}{3}$.

Now for our theorem ...

Theorem 9.13.1. (Bayes' Theorem)

$$p(B \mid A) = \frac{1}{p(A \mid B) \cdot p(B) + p(A \mid \overline{B}) \cdot p(\overline{B})} \cdot p(A \mid B) \cdot p(B)$$

Let's prove this theorem. Recall that

$$p(B \mid A) \cdot p(A) = p(A \mid B) \cdot p(B)$$

Therefore

$$p(B \mid A) = \frac{1}{p(A)} \cdot p(A \mid B) \cdot p(B)$$

Note that we just need to prove

$$p(A) = p(A \mid B) \cdot p(B) + p(A \mid \overline{B}) \cdot p(\overline{B})$$

The right hand side is

$$\begin{aligned} p(A \mid B)p(B) + p(A \mid \overline{B})p(\overline{B}) &= \frac{p(A \cap B)}{p(B)}p(B) + \frac{p(A \cap \overline{B})}{p(\overline{B})}p(\overline{B}) \\ &= p(A \cap B) + p(A \cap \overline{B}) \end{aligned}$$

The two events $A \cap B$ and $A \cap \overline{B}$ are clearly disjoint, and therefore

$$p(A \cap B) + p(A \cap \overline{B}) = p(A)$$

and we're done! □

QUESTION: Which is better

$$p(B \mid A) = \frac{1}{p(A \mid B) \cdot p(B) + p(A \mid \overline{B}) \cdot p(\overline{B})} \cdot p(A \mid B) \cdot p(B)$$

or

$$p(B \mid A) = \frac{1}{p(A \cap B) + p(A \cap \overline{B})} \cdot p(A \mid B) \cdot p(B)$$

In other words this is better:

$$p(A \mid B) \cdot p(B) + p(A \mid \overline{B}) \cdot p(\overline{B})$$

or

$$p(A \cap B) + p(A \cap \overline{B})$$

Bayes' theorem can be generalized in the following way. The basic Bayes' theorem say this:

$$p(B \mid A) = \frac{1}{p(A \mid B) \cdot p(B) + p(A \mid \overline{B}) \cdot p(\overline{B})} \cdot p(A \mid B) \cdot p(B)$$

This comes from

$$p(B \mid A)p(A) = p(A \mid B)p(B)$$

which gives us

$$p(B \mid A) = \frac{1}{p(A)} \cdot p(A \mid B)p(B)$$

The term $p(A)$ is then computed using

$$p(A) = p(A \cap B) + p(A \cap \overline{B})$$

Hence it's easy to see that if instead of B and $\overline{B}$, I have a collection of events $B_0, ..., B_{n-1}$ such that $B_i \cap B_j = \emptyset$ for $i \neq j$ and $\cup_{i=0}^{n-1} B_i = S$, then

$$p(B_i \mid A)p(A) = \frac{1}{p(A)} \cdot p(A \mid B_i)p(B_i)$$

and

$$p(A) = \sum_{i=0}^{n-1} p(A \cap B_i)$$

Putting all the above together, I get

$$p(B_i \mid A) = \frac{1}{\sum_{i=0}^{n-1} p(A \cap B_i)} \cdot p(A \mid B_i)p(B_i)$$

Now from

$$p(A \mid B_i) = \frac{p(A \cap B_i)}{p(B_i)}$$

I get

$$p(A \mid B_i)p(B_i) = p(A \cap B_i)$$

Therefore

$$p(B_i \mid A) = \frac{1}{\sum_{i=0}^{n-1} p(A \cap B_i)} \cdot p(A \mid B_i)p(B_i) = \frac{1}{\sum_{i=0}^{n-1} p(A \mid B_i)p(B_i)} \cdot p(A \mid B_i)p(B_i)$$

File: discrete-probability/probabilistic-algorithms.tex

9.14 Randomized/probabilistic algorithms

A **randomized algorithm** or **probabilistic algorithm** is an algorithm that uses randomization.

randomized algorithm
probabilistic algorithm

There are two types of randomized/probabilistic algoroithms.

A **Las Vegas** randomized algorithm uses randomization of input. The output is correct. What is changed, because of the randomization of input, is the expected runtime. An example is randomized quicksort where the choice is pivot is randomized.

Las Vegas

A **Monte Carlo** randomized algorithm might produce an incorrect result, but with a very small probability. An example is the Miller-Rabin primality testing algorithm. In the case of primality testing, several rounds of Miller-Rabin will lower the probability of incorrectness.

Monte Carlo

Now let's consider quicksort. Let $T(n)$ be the runtime. Once a pivot is selected, recall that the pivot partitions the subarray the quicksort is working on into two parts (the left and right partitions), but in general you do not know the sizes of the two partitions. Therefore the runtime looks like this:

$$T(n) = T(0) + T(n-1) + An + B$$

or

$$T(n) = T(1) + T(n-2) + An + B$$

or

$$T(n) = T(2) + T(n-3) + An + B$$

or

$$\ldots$$

or

$$T(n) = T(n-1) + T(0) + An + B$$

depending on how the pivot partitions. For instance for the first line above, the pivot lands in index 0, and therefore the left partition has size 0 and the right partition has size $n-1$. Let's assume all the above cases are equally likely. Adding them up I get

$$nT(n) = 2 \cdot \sum_{i=0}^{n-1} T(i) + n(An + B)$$

and therefore

$$T(n) = \frac{2}{n}\sum_{i=0}^{n-1} T(i) + An + B$$

Note that technically the $T(n)$ here is the average $T(n)$.

Now I want to show that the above average runtime is $O(n \lg n)$.

Suppose $T(k) \leq k \lg k$ for $1 \leq k < n$. Then

$$\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n-1} T(i) &\leq \frac{2}{n}\sum_{i=1}^{n-1} i \lg i \\
&\leq \frac{2}{n}\sum_{i=1}^{n-1} n \lg i \\
&\leq 2\sum_{i=1}^{n-1} \lg i \\
&\leq 2\int_1^n \lg x \; dx \\
&\leq 2C\int_1^n x \ln x \; dx \qquad \text{change of base}
\end{aligned}$$

9.14.1 Average runtime of quicksort using probability

Assume `x[0..n-1]` values are distinct and I'm going to perform quicksort on it. I'm going to assume that the pivot selection is random.

I'm using the partition method mentioned in my notes where the pivot is placed at the beginning of the subarray to be sorted.

I'll compute the average runtime this way: I'm going to count the number of comparisons throughout the *whole* quicksort process, instead of going through the stages of recursion. Again, as before, the above statement means "I'm going to count the *average* number of comparison".

Let X be number of comparisons during quicksort. So I'm aiming for $\mathrm{E}[X]$. Let's break it up. Let X_{ab} be the number of comparison between `a` and `b` where `a` and `b` are values in the array `x[0..n - 1]` that I'm sorting. Then

$$X = \sum_{a,b \in \texttt{x[0..n-1]}} X_{ab}$$

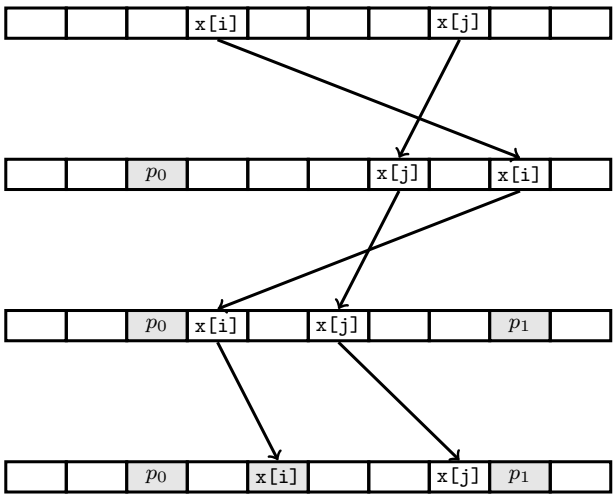
and therefore

$$\mathrm{E}[X] = \sum_{\texttt{a,b} \in \texttt{x[0..n-1]}} \mathrm{E}\left[X_{\texttt{ab}}\right]$$

Note that `a` and `b` is compared at most once. Why? Because values are compared *against a pivot* and once a value is chosen as a pivot, you notice that work is never done on that pivot anymore – because the pivot has found its right place during the partitioning algorithm. Of course it's also possible that `a` and `b` are never chosen as pivot. In that case the number of comparisong between `a` and `b` is 0. Therefore $X_{\texttt{ab}} = 1$ or 0. In other words $X_{\texttt{ab}}$ is an indicator random variable. Therefore

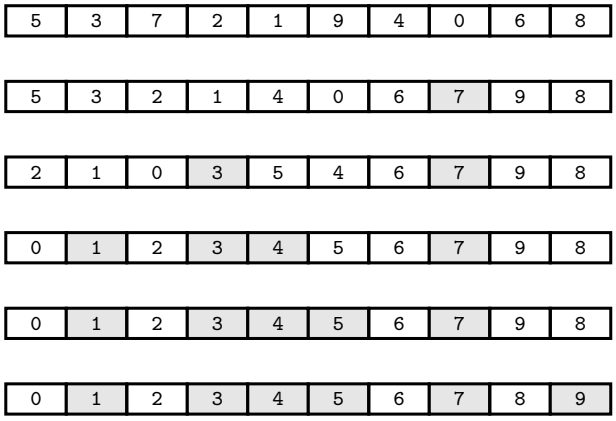
$$\begin{aligned} E[X] &= \sum_{\texttt{a,b} \in \texttt{x[0..n - 1]}} E\left[X_{\texttt{ab}}\right] \\ &= \sum_{\texttt{a,b} \in \texttt{x[0..n - 1]}} \Pr\left[X_{\texttt{ab}} = 1\right] \\ &= \sum_{\texttt{a,b} \in \texttt{x[0..n - 1]}} (\text{probability that } \texttt{a} \text{ and } \texttt{b} \text{ are compared}) \end{aligned}$$

Here's a diagram showing a sample execution of quicksort on `x[0..n - 1]`, focusing on two elements `x[i]` and `x[j]` where `x[i]` is ultimately chosen as a pivot:



After that `x[j]` might change it's position. But it won't be compared against `x[i]` anymore. The cells in grey are pivots after the partitioning steps and therefore will not take part in future partitioning steps.

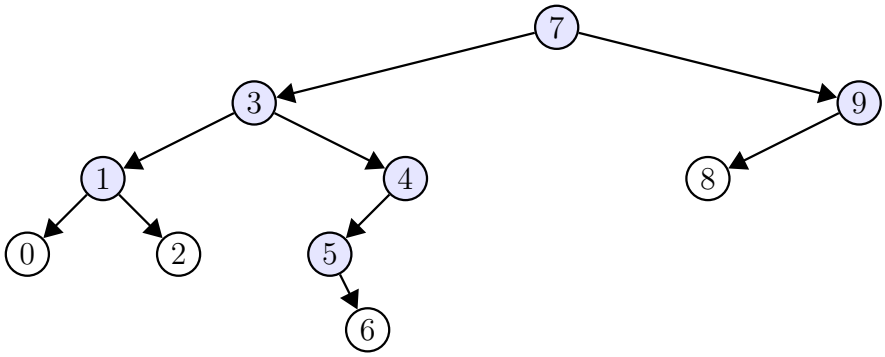
Here's the diagram of a more complete quicksort on an array of size 10. I'm drawing the array at the end of each pass, which includes the pivot selection and the partitioning step. The pivots are shaded.



If you look at

where the pivots are shaded, note that `0` is compared against the nodes on the path from the root `7` to `0`, not including `0`.

Here's the binary tree corresponding to the quicksort where the parent nodes are the pivots:



The nodes which are shaded are the pivots. This tree diagram makes it easier to see that `0` is compared against `1`, `3`, and `7`.

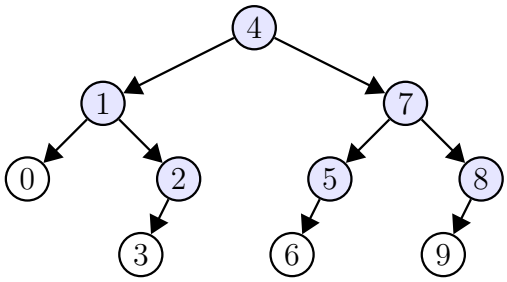
Altogether, the total number of comparisons is the sum of the lengths of all paths in the above tree. Therefore `0` is compared against `1`, `3`, and `7`. Altogether X for this case is

- Number of values compared against 7: 9
- Number of values compared against 3: 6

- Number of values compared against 1: 2
- Number of values compared against 4: 2
- Number of values compared against 5: 1
- Number of values compared against 9: 1

Therefore the total number of comparisons is 21.

Here's a tree where pivots are chosen so that the tree is height-balanced:



The number of comparisons is 19.

Analyzing

$$\mathrm{E}[X] = \sum_{\texttt{a,b}\in\texttt{x[0..n - 1]}} \text{(probability that } \texttt{a} \text{ and } \texttt{b} \text{ are compared)}$$

where `a` is (say) `x[i]` and `b` is `x[j]` where `x[0..n - 1]` is before the execution of quicksort is difficult. So (and this is important), I'm going to let `x[0..n - 1]` be the array *after* sorting is done. Note that

$$\mathrm{E}[X] = \sum_{\texttt{a,b}\in\texttt{x[0..n - 1]}} \text{(probability that } \texttt{a} \text{ and } \texttt{b} \text{ are compared)}$$

does not depend on whether where `a` = `x[i]` and `b` = `x[j]` are values before or after quicksort. Remember what I said earlier: for `a` and `b` to be compared, either `a` or `b` has to be chosen as pivot. Also, recall that I'm using `x[0..n - 1]` to denote our array after it is sorted. Here's the important observation:

Fact. `x[i]`, `x[j]` are compared iff the first value in `x[i..j]` to be chosen as pivot is either `x[i]` or `x[j]`.

($\Longleftarrow$): Easy. If `x[i]` is the first in `x[i..j]` to be selected as pivot, then `x[j]` will be compared against `x[i]`. Likewise if `x[j]` in the first in `[i..j]` to be selected as pivot, then `x[i]` will be compared against `x[j]`.

($\Longrightarrow$): Assume some `x[k]` where `x[i]` $<$ `x[k]` $<$ `x[j]` is first selected as pivot among the values `x[i..j]`. Then `x[i]` and `[j]` would have landed in

different partitions after `x[k]` separates them. After that, they obviously will not be able to compare with each other.

Hence, the probability that `x[i]` and `x[j]` are compared is the probability that the first value in `x[i..j]` to be chosen as pivot is either `x[i]` or `x[j]`. There are $j-i+1$ values in `x[i..j]`. Assuming equal likelihood of choosing any value `x[i..j]` to be a pivot,

$$\begin{aligned}&\text{probability that } \texttt{x[i]} \text{ or } \texttt{x[j]} \text{ is first pivot chosen among } \texttt{x[i..j]}\\ &= \frac{2}{j-i+1}\end{aligned}$$

For instance the probability that `x[i]` or `x[i + 1]` is first pivot chosen as pivot along `x[i..i + 1]` is $2/(i+1-i+1) = 2/2 = 1$. Hence

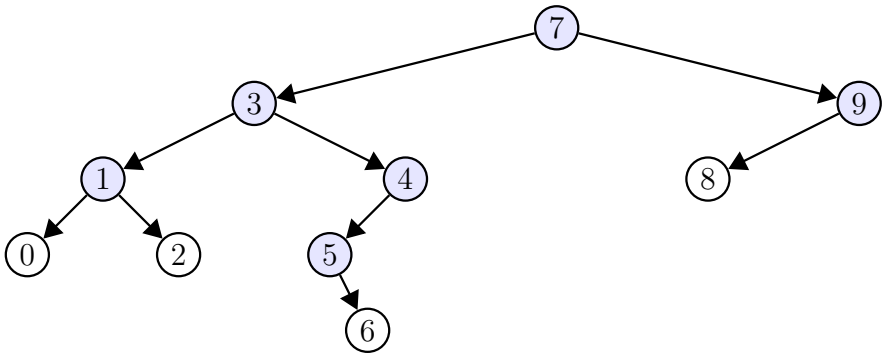
$$\begin{aligned}\mathrm{E}[X] &= \sum_{0\le i<j\le n-1} \frac{2}{j-i+1}\\ &= 2\sum_{i=0}^{n-2}\sum_{j=i+1}^{n-1}\frac{1}{j-i+1}\\ &= 2\sum_{i=0}^{n-2}\sum_{k=2}^{n-i}\frac{1}{k} && (\text{let } k = j-i+1)\\ &\le 2\sum_{i=0}^{n-2}\sum_{k=1}^{n}\frac{1}{k}\\ &= 2\sum_{i=0}^{n-2} H_n\\ &= 2(n-1)H_n\\ &\le 2nH_n\end{aligned}$$

Since $\lim_{n\to\infty}(H_n - \ln n) = \gamma = 0.577...$, there is some N such that for $n \ge N$, $H_n \le \ln n + 1$. Hence $H_n = O(\ln n) = O(\lg n)$. Therefore

$$\mathrm{E}[X] \le 2nH_n = O(n \lg n)$$

To make sure you understand the proof above, let me look at the quicksort from above:

and its tree:



The number of comparison involving `0` is 3. The *expected* number of comparisons involving `0` (for all quicksorts) is given by

i	j	`x[j]` compared $= 2/(i+j-1)$	X_{ij}
0	1	$2/(1-0+1) = 2/2 = 1$	1
0	2	$2/(2-0+1) = 2/3 = 0.6666...$	0
0	3	$2/(3-0+1) = 2/4 = 0.5$	1
0	4	$2/(4-0+1) = 2/5 = 0.4$	0
0	5	$2/(5-0+1) = 2/6 = 0.3333...$	0
0	6	$2/(6-0+1) = 2/7 = 0.2857...$	0
0	7	$2/(7-0+1) = 2/8 = 0.25$	1
0	8	$2/(8-0+1) = 2/9 = 0.2222...$	0
0	9	$2/(9-0+1) = 2/10 = 0.2$	0
		Sum $= 3.8579...$	Sum $= 3$

There are 3 comparisons involving 0 and the average number of comparisons (over all possible quicksort) is 3.8579... For $n = 10$, the expected number of comparisons is

$$\mathrm{E}[X] = 2\sum_{i=0}^{10-2}\sum_{k=2}^{10-i}\frac{1}{k} = 32.9$$

Exercise 9.14.1. So you have decided to be biased when you choose your pivot: You prefer to choose value in the middle of the subarray you are quicksorting. In other words while quicksorting `x[i..j]`, you prefer `x[(i + j)/2]` (the division is integer division. You do want the other values to have a chance of being picked with equal likelyhood adding up to 0.5. To be more precise, you choose `x[(i + j)/2]` with a probability of 0.5 and the probability of choosing `x[k]` for `k` $\neq$ `(i + j)/2` is `(0.5)/(j - i)`. Write a quicksort function that selects a pivot randomly (with uniform pdf) and then another one that selects

that prefers the value in the middle as above. Run your program and collect data on the average number of comparisons for different n (the size of the array to be sorted). What is the expected number of comparisons for the quicksort that prefers the value in the middle? □

9.14.2 Average runtime without probability theory

Another way to compute the average runtime is to derive a closed form for $T(n)$ where

$$T(n) = \frac{2}{n}\sum_{i=0}^{n-1} T(i) + An + B$$

for sufficiently large n. Here, sufficiently can be, say, $n > 1$, using $n = 0, 1$ as base case since there's no reason to use quicksort if the array has size 0 or 1.

There are many ways to proceed from this point.

9.14.3 Method A: Induction

The problem with induction is always that you need to know what you need to prove. In this case, I have to somehow suspect that $T(n) = O(n \lg n)$. This can be achieved by doing some timings. So say I claim that

$$T(n) = O(n \lg n)$$

i.e., there is some N and some C such that

$$T(n) \leq Cn \lg n$$

for $n \geq N$.

Then assuming that

$$T(k) \leq Ck \lg k$$

for $k = N, N+1, ..., n-1$, I need to prove that

$$T(n) \leq Cn \lg n$$

I get

$$\begin{aligned}
T(n) &= \frac{2}{n}\sum_{i=0}^{n-1} T(i) + An + B \\
&\leq \frac{2}{n}\left(T(0) + B + T(1) + A + B\right) + \frac{2}{n}\sum_{k=2}^{n-1} T(k) + An + B \\
&\leq \frac{2}{n}\left(T(0) + T(1) + A + 2B\right) + \frac{2}{n}\sum_{k=2}^{n-1} Ck\lg k + An + B \\
&\leq \frac{2C}{n}\sum_{k=2}^{n-1} k\lg k + \frac{2}{n}\left(T(0) + T(1) + A + 2B\right) + An + B
\end{aligned}$$

The sum can be bounded this way (converting the sum to an area computation):

$$\sum_{k=2}^{n-1} k\lg k \leq \int_2^n x\lg x\,dx$$

Fortunately we do have a relevant integration formula for this case which you can find in a calculus book:

$$\int x\ln x\;dx = \frac{1}{2}x^2\ln x - \frac{1}{4}x^2 + C$$

Or you can derive the formula using integration-by-parts:

$$\begin{aligned}
\int x\ln x\;dx &= \int \ln x\;d\left(\frac{1}{2}x^2\right) \\
&= \frac{1}{2}x^2\ln x - \int \frac{1}{2}x^2\;d(\ln x) \\
&= \frac{1}{2}x^2\ln x - \frac{1}{2}\int x^2\frac{1}{x}\;dx \\
&= \frac{1}{2}x^2\ln x - \frac{1}{2}\int x\;dx \\
&= \frac{1}{2}x^2\ln x - \frac{1}{2}\cdot\frac{1}{2}x^2 + C \\
&= \frac{1}{2}x^2\ln x - \frac{1}{4}x^2 + C
\end{aligned}$$

I do have to change the log to base 2:

$$\int x\lg x\;dx = \frac{1}{2}x^2\lg x - \frac{\alpha}{4}x^2 + C$$

where $\alpha = 1/\log_e 2$. Therefore I now have

$$\begin{aligned}
\sum_{k=2}^{n-1} k \lg k &\leq \int_2^n x \lg x \, dx \\
&= \left(\frac{1}{2}x^2 \ln x - \frac{1}{4}x^2\right)\Bigg|_{x=2}^{x=n} \\
&= \left(\frac{1}{2}n^2 \ln n - \frac{1}{4}n^2\right) - \left(\frac{1}{2}2^2 \lg 2 - \frac{\alpha}{4}2^2\right) \\
&= \left(\frac{1}{2}n^2 \lg n - \frac{1}{4}n^2\right) - (2 \lg 2 - \alpha)
\end{aligned}$$

Going back to $T(n)$:

$$\begin{aligned}
T(n) &\leq \frac{2C}{n}\sum_{k=2}^{n-1} k \lg k + \frac{2}{n}\left(T(0) + T(1) + A + 2B\right) + An + B \\
&\leq \frac{2C}{n}\left(\left(\frac{1}{2}n^2 \lg n - \frac{1}{4}n^2\right) - (2 \lg 2 - \alpha)\right) + \frac{2}{n}\left(T(0) + T(1) + A + 2B\right) + An + B \\
&\leq Cn \lg n + \frac{2C}{n}\left(\left(-\frac{1}{4}n^2\right) - (2 \lg 2 - \alpha)\right) + \frac{2}{n}\left(T(0) + T(1) + A + 2B\right) + An + B
\end{aligned}$$

The RHS is

$$\leq Cn \lg n$$

if

$$\frac{2C}{n}\left(\left(-\frac{1}{4}n^2\right) - (2 \lg 2 - \alpha)\right) + \frac{2}{n}\left(T(0) + T(1) + A + 2B\right) + An + B$$

is ≤ 0. In other words, I need

$$\frac{2}{n}\left(T(0) + T(1) + A + 2B\right) + An + B \leq \frac{2C}{n}\left(\left(\frac{1}{4}n^2\right) + (2 \lg 2 - \alpha)\right)$$

i.e.,

$$2\left(T(0) + T(1) + A + 2B\right) + An^2 + Bn \leq 2C\left(\left(\frac{1}{4}n^2\right) + (2 \lg 2 - \alpha)\right)$$

i.e.,

$$0 \leq \left(\frac{C}{2} - A\right)n^2 - Bn + \beta$$

where $\beta = 2C(2\lg 2 - \alpha) - 2\left(T(0) + T(1) + A + 2B\right)$. If I choose C such that $C/2 - A > 0$, then the parabola function above $(C/2 - A)n^2 - Bn + \beta$ will concave up and will be > 0 for $n > N_0$ for some N_0.

Altogether, there is some N and some C such that

$$T(n) \leq Cn \lg n$$

for $n \geq N$.

9.14.4 Method C: Direct derivation

In the previous section, I have to make a guess that $T(n) = O(n \lg n)$. What if I want to derive the runtime without guessing?

From the previous section, I have

$$T(n) = \frac{2}{n} \sum_{i=0}^{n-1} T(i) + An + B$$

for $n > 1$.

Suppose I want to derive the a closed-form for $T(n)$. I'm in trouble because the recursion is not a linear degree 2 recursion and it does not fit the form of the master theorem. Not only is it not degree 2, in fact, it's degree n – i.e., the degree is not fixed.

Don't panic. First do this:

$$nT(n) = 2 \sum_{i=0}^{n-1} T(i) + An^2 + Bn$$

Therefore, with n replaced by $n - 1$, I get

$$(n-1)T(n-1) = 2 \sum_{i=0}^{n-2} T(i) + A(n-1)^2 + B(n-1)$$

Subtracting:

$$nT(n) - (n-1)T(n-1) = 2T(n-1) + A(n^2 - (n-1)^2) + B(n - (n-1))$$

So what? ... well I have removed the $\sum$ so that I now have

$$\begin{aligned} nT(n) &= (n+1)T(n-1) + A(2n+1) + B \\ &= (n+1)T(n-1) + 2An + (A+B) \\ &= (n+1)T(n-1) + A'n + B' \\ \therefore\ T(n) &= \left(1 + \frac{1}{n}\right) T(n-1) + A' + \frac{B'}{n} \end{aligned}$$

which is a degree 1 recurrence, although unfortunately it's not linear. There are many ways to solve this ...

9.14.5 Method C1: telescoping trick

From the above, relation

$$T(n) = \left(\frac{n+1}{n}\right) T(n-1) + A' + \frac{B'}{n}$$

I get

$$\frac{1}{n+1}T(n) = \frac{1}{n}T(n-1) + \frac{A'}{n+1} + \frac{B'}{n(n+1)}$$

and therefore

$$\frac{1}{n+1}T(n) - \frac{1}{n}T(n-1) = \frac{A'}{n} + \frac{B'}{(n-1)n}$$

Therefore I get

$$\begin{aligned} \frac{1}{n+1}T(n-0) - \frac{1}{n-0}T(n-1) &= \frac{A'}{n+1} + \frac{B'}{(n-0)(n+1)} \\ \frac{1}{n+0}T(n-1) - \frac{1}{n-1}T(n-2) &= \frac{A'}{n+0} + \frac{B'}{(n-1)(n+0)} \\ \frac{1}{n+(-1)}T(n-2) - \frac{1}{n-2}T(n-3) &= \frac{A'}{n+(-1)} + \frac{B'}{(n-2)(n+(-1))} \\ \frac{1}{n+(-2)}T(n-3) - \frac{1}{n-3}T(n-4) &= \frac{A'}{n+(-2)} + \frac{B'}{(n-3)(n+(-2))} \\ &\vdots \end{aligned}$$

Adding these, I get

$$\frac{1}{n+1}T(n-0) - \frac{1}{2}T(1) = A' \sum_{k=3}^{n+1} \frac{1}{k} + B' \sum_{k=2}^{n} \frac{1}{k(k+1)}$$

i.e.,

$$T(n) = A'(n+1)\sum_{k=3}^{n+1}\frac{1}{k} + \frac{n+1}{2}T(1) + B'(n+1)\sum_{k=2}^{n}\frac{1}{k(k+1)}$$

and you get $O(n \lg n)$.

The telescoping trick above depends on the recursion on being degree 1 (but not necessarily linear) of the form

$$f(n)T(n) = f(n-1)T(n-1) + g(n)$$

for $n \geq n_0$ and therefore

$$f(n)T(n) - f(n-1)T(n-1) = g(n)$$

Therefore

$$\begin{aligned} f(n)T(n) - f(n-1)T(n-1) &= g(n) \\ f(n-1)T(n-1) - f(n-2)T(n-2) &= g(n-1) \\ f(n-2)T(n-2) - f(n-3)T(n-3) &= g(n-2) \end{aligned}$$

etc. Therefore

$$f(n)T(n) - f(n_0-1)T(n_0-1) = g(n) + \cdots + g(n_0)$$

and hence

$$T(n) = \frac{g(n) + \cdots + g(n_0)}{f(n)} + \frac{C}{f(n)}$$

where $C = f(n_0-1)T(n_0-1)$.

9.14.6 Method C2: generating functions

From the previous section I have

$$T(n) = \left(1 + \frac{1}{n}\right)T(n-1) + A' + \frac{B'}{n}$$

for $n > 0$. Not all degree 1 non-linear can be put into a suitable form for telescoping. A more general method is the method of generating functions. Let

$$f(x) = \sum_{n=0}^{\infty} T(n)x^n$$

Exercise 9.14.2. Using the fact that

$$\int_{t=0}^{t=x} \frac{1}{n} t^{n-1} \, dt = \frac{1}{n} x^n$$

show that $f(x)$ satifies an equation of the form

$$(1-x)f(x) - \int_0^x f(t) \, dt = T(0) + C \int_0^x \frac{1}{1-t} \, dt + \frac{Dx}{1-x}$$

where C and D are constants. □

Exercise 9.14.3. Take the derivative (wrt x) for the equation in the previous exercise and your $f(x)$ will satisfy a differential equation of the form:

$$y' + p(x)y = q(x)y^n$$

where the right is (fortunately) simple: it's $q(x)$ instead of $q(x)y^n$, i.e., $n = 0$. The above is called a Bernoulli's differential equation. □

Exercise 9.14.4. Following the standard way to solve a Bernoulli differential equation, let

$$m(x) = e^{\int p(x) \, dx}$$

Then $f(x)$ is given by

$$f(x) = \frac{\int m(x)q(x) \, dx + E}{m(x)}$$

where E is a constant. □

Exercise 9.14.5. Write $f(x)$ from the previous exercise as a power series $\sum_{n=0} a_n x^n$. The n–th coefficient a_n is $T(n)$. You should see that $T(n)$ is $\Theta(n \lg n)$. In fact you should be able to get an exact closed form for $T(n)$. □

Index