Lecture 3: Random Variables and Expectation

Rajat Mittal

IIT Kanpur

We have looked at the basics of probability (and conditional probability). For a random experiment, three most important things were sample space, σ-field and probability distribution function. More often than not, σ-field is the power set of the sample space and not explicitly specified. From our definition, sample space is just a set, elements could be anything. They can be colors, persons, faces of dice, numbers and so on.

Many a times, we are interested in a *numerical value* associated to the outcomes of the experiment (sample space). For example, number of heads in a sequence of tosses, pay out of a lottery, number of casualties after a Hurricane etc. These functions, which assign a numerical value to the outcomes of the experiment, are called *random variables.* In this lecture, we will look at random variables and their properties.

1 Random variable

Given the sample space Ω of an experiment, a random variable is a function $X : \Omega \to \mathbb{R}$. That means, a random variable assigns a real value $X(\omega)$ to every element ω of the sample space.

Note 1. In general, the range of a random variable X need not be $\mathbb{R}$. It could be any other set with more structure (like real numbers are ordered; they can be added, multiplied etc.). When range is real numbers, such random variables are called *real valued.* In this course, we will mostly be interested in real valued random variables, and we will simply call them random variables.

If the range of X is countable then X is called a discrete random variable. We will initially focus on discrete random variables.

A random variable gives rise to events of the type $X = x$, which is associated with a subset $X^{-1}(x)$ of the sample space Ω. Given a probability function P on Ω, it can be naturally extended to the probability of the random variable,

$$P_X(x) := P(X = x) = \sum_{\omega : X(\omega) = x} P(\omega).$$

This is called the *probability mass function* of a random variable. Let's look at some examples of random variables and their probability mass function.

- Suppose you toss a fair coin 10 times. The sample space Ω is the set of all sequences of length 10 made up with H, T. Define a random variable $X : \Omega \to \mathbb{R}$ to be the number of heads in the sequence. That is, $X(\omega)$ is the number of H's in ω.

  Exercise 1. Show that the probability of getting a sequence with k heads for a length 10 sequence is,

  $$\binom{10}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{10-k}.$$

  The probability mass function of the random variable for k between 0 and 10 becomes,

  $$P(X = k) = \left(\frac{1}{2}\right)^{10} \binom{10}{k}.$$

  Exercise 2. Generalize the probability mass function if the probability of obtaining head is p.

- For an experiment, we ask the birthday's of students in a class one by one (picking the next student randomly). We stop as soon as we find two people with matching birthday. What is the probability mass function for the random variable X which counts the number of students queried? What is the probability that we queried k people?

  If we had to query k people, then the first $k-1$ birthdays are distinct and the last one matches at least one of the first $k-1$. First birthday will not match with anyone before that with probability 1. Second person's birthday will not match with probability 364/365 and so on. The birthday of the last person needs to match, happens with probability $(k-1)/365$.

  Hence,

  $$Pr(X = k) = \frac{(k-1)}{365^k}(k-1)!\binom{365}{k-1}.$$

  Exercise 3. Calculate these numbers using a calculator for k up to 23.

- In a set of 1000 balls, 150 balls have some defect. Say, we choose 50 balls and inspect, then let X be the random variable which denotes the number of defected balls found.

  The probability mass function of X is non-zero for $x = 1$ to 50. It is given by,

  $$Pr(X = k) = \frac{\binom{150}{k}\binom{850}{50-k}}{\binom{1000}{50}}.$$

  Exercise 4. What if we take out the balls one by one and return them to the set before the next pick.

Functions of random variables: We can have many random variables for a random experiment. To start with, we can take any function $g : \mathbb{R} \to \mathbb{R}$ and a random variable X to create another random variable $Y = g(X)$ on the same sample space.

Exercise 5. Notice that Y can be thought of as a function from Ω to $\mathbb{R}$, how? What will be the range of Y in terms of range of X?

It is not difficult to come up with an expression for the probability mass function (PMF) of the new random variable Y.

$$\begin{aligned} P_Y(y) =& P(Y = y) \\ =& P(g(X) = y) \\ =& \sum_{x:g(x)=y} P(X = x) \\ =& \sum_{x:g(x)=y} P_X(x) \end{aligned} \tag{1}$$

Let us take an example, suppose X is the random variable which counts the number of heads in a sequence of 10 coin tosses (where probability of head is p), remember

$$P(X = k) = p^k(1-p)^{10-k}\binom{10}{k}.$$

Define $Y = X^2$, then the PMF of Y would be

$$P(Y = k) = p^{\sqrt{k}}(1-p)^{10-\sqrt{k}}\binom{10}{\sqrt{k}},$$

if $0 \leq k \leq 100$ is a perfect square and 0 otherwise.

On a similar note, we can have functions of multiple random variables, like $X+Y, XY$ etc. Let us consider $X+Y$, where X, Y are two random variables on the same sample space. What would be the PMF of $X+Y$, given the PMF of X, Y. Intuitively,

$$P(X+Y=k) = \sum_{x,y:x+y=k} P(X=x, Y=y).$$

What should be $P(X=x, Y=y)$ in terms of $P(X=x)$ and $P(Y=y)$? To illustrate the difficulty, consider that we roll a dice twice and X be the outcome on first roll and Y be the outcome on the second roll. You would guess that $P(X=x, Y=y) = P(X=x)P(Y=y)$. It is good for this case, but what if $X=Y$, would you want to still say $P(X=x, Y=y) = P(X=x)P(Y=y)$. Figuring out $P(X=x, Y=y)$ in terms of $P(X=x)$ and $P(Y=y)$ is a deep question and we will be spend a lot of time in the next lecture discussing this question (conditional probability and independence).

For now, let us skip this issue and just define *joint probability mass function*. The joint probability mass function of X, Y is defined to be

$$P_{X,Y}(x,y) := P(X=x \text{ and } Y=y).$$

Notice that a valid joint distribution will satisfy $\sum_{x,y} P(X=x, Y=y) = 1$. If individual PMF's are given then $P(X=x) = \sum_y P(X=x, Y=y)$ and a similar equation for $P(Y=y)$ should also be satisfied. Otherwise, in case only joint PMF is given, it will tell us the individual PMF by these equations.

If we want to compute the PMF of $X+Y$ (or any function of X, Y), we need the joint probability mass function of X, Y.

2 Expectation

Remember that random variables were introduced because of our interest in numerical values associated with the set of outcomes. These real values allow us to talk about their cumulative behavior (like average and deviation). For example, if you get 1 rs. for a heads and 2 rs. for tails, you might want to estimate your earnings in a sequence of 10 tosses.

The most basic cumulative quantity is called the *expectation* of a random variable. It is easy to define if all outcomes are equally likely (known as *average*). In this case, if $\Omega = \{\omega_1, \cdots, \omega_n\}$, we would expect to get the average

$$\left(\frac{X(\omega_1) + X(\omega_2) + \cdots + X(\omega_n)}{n}\right).$$

Taking this idea further, the expected value of a random variable X is defined as,

$$E[X] := \sum_{x \in \mathbb{R}} Pr(X(\omega) = x)x.$$

When random variables are discrete, range of X might be much smaller than $\mathbb{R}$ (by range of X, we mean values attained by X with non-zero probability). Let R denote the range of X, simply

$$E[X] := \sum_{x \in R} Pr(X(\omega) = x)x.$$

It is a common misinterpretation, probably because of name, that X will attain value $E[X]$ with high probability. It is easy to construct cases where $E[X]$ might not even be in R.

Exercise 6. Construct a random variable such that $E[X]$ is not in R.

The correct intuition is: if we independently repeat the experiment multiple times, then with high probability the average outcome will be close to expectation. This intuition will be formalized in later parts of the course.

To start, solve a simple problem on expectation.

Exercise 7. In a probabilistic experiment, you get 100 Rs. every time an odd number shows up on a dice. You loose 100 Rs. every time an even number shows up. What is your expected earning.

How about some more examples of expectation?

- Your friend is ready to give you 100 Rs. if on a throw of a dice, an odd prime turns up. What amount can you give him if the number is not an odd prime?

  Exercise 8. What is the random variable in this case?

  You cannot always be certain to win this game if you bet any positive amount. We can define the bet to be profitable if your expected profit is greater than zero (that means, at least you win when the experiment is repeated multiple times).
  If you bet x Rs., then the expected earning should be greater than zero, $1/3 \times 100 + 2/3 \times (-x) \geq 0$. So you can agree to pay any amount less than 50. This example suggests that in a fair bet, the expected profit/loss should be zero.

  Exercise 9. Suppose the expected value of a random variable X is zero. What is the expected value of $-X$? What is the expected value of αX for a constant α?

- You toss a coin till you get head (biased coin with head probability p). What is the expected number of tosses? Clearly, the random variable is the number of tosses.

  Exercise 10. What was the sample space though? Show that $Pr(X = k) = (1-p)p^{k-1}$.

  The expected number of tosses is,
  $$E[X] = \sum_k k(1-p)p^{k-1}.$$
  Let $S = \sum_k k(1-p)p^{k-1}$, then $pS = (1-p)\sum_k kp^k$. Subtracting these two equations tells us that $S = 1/p$.
- Your friend asks you to bet on the rise/fall of stock market. Both of you put 100 Rs. in a pot and guess. If the guesses are same then both get 100 Rs., otherwise the one with the correct guess gets all the money. Is this bet fair, assuming that none of you have any clue about share market?
  Suddenly, your friend wants to include her brother's share also. She proposes to put 200 Rs. in the pot and will guess for both her and her brother. Should you take the bet?
  Suppose you friend puts opposite guesses for her and her brother all the time. If you guess is correct then you will get 150 Rs. and if you incorrectly guess then you get 0 Rs.. The expected value of your pot earning is 75 Rs., which is less than the amount you put in the pot. So it is not an advisable bet.

One point of caution is that expectation need not be defined all the time. Let X be a random variable such that $Pr(X = k) = \frac{6}{\pi^2 k^2}$. It can be shown that $\sum_k Pr(X = k) = 1$ but $\sum_k Pr(X = k)k$ is not convergent.

We saw that if X is a random variable then so is $Y = g(X)$ where g is any function from $\mathbb{R}$ to $\mathbb{R}$. Then,
$$E[Y] = E[g(X)] = \sum_{x \in \mathbb{R}} g(x)Pr(X = x).$$

Note 2. We need to assume that $\sum_x |g(x)|Pr(X = x)$ converges.

2.1 Linearity of expectation

One of the most important property of expectation is that it is linear. This might seem like a straightforward property of expectation, but its applications are huge and hence deserves a separate section.

Given two random variables X and Y on the same sample space Ω,
$$E[X + Y] = E[X] + E[Y].$$

Note 3. Here say $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$, then random variable $X + Y : \Omega \to \mathbb{R}$ is defined as $X + Y(\omega) = X(\omega) + Y(\omega),\ \forall \omega \in \Omega$.

This property is known as *linearity of expectation.* To get some intuition, think about averages. If you know the average marks of students in each subject, can you find the average total marks (sum of marks) over all students.

Proof. The expectation $E[X + Y]$ is given using the probability mass function $Pr(X = x, Y = y)$.

$$\begin{aligned} E[X+Y] &= \sum_{x,y} (x+y) Pr(X = x, Y = y) \\ &= \sum_{x,y} x Pr(X = x, Y = y) + y Pr(X = x, Y = y) \\ &= \sum_{x} x \sum_{y} Pr(X = x, Y = y) + \sum_{y} y \sum_{x} Pr(X = x, Y = y) \\ &= \sum_{x} x Pr(X = x) + \sum_{y} y Pr(Y = y) \\ &= E[X] + E[Y] \end{aligned} \tag{2}$$

□

The linearity of expectation can be extended to more than two events using induction.

Exercise 11. What will be the statement?

The property seems almost obvious and even the proof is pretty straightforward. In spite of that, linearity of expectation is one of the very important tools and is used very frequently. One of the reason being, nothing is assumed about the relationship between X and Y. The random variables need not depend on each other in any way, still linearity of expectation holds true. In contrast, $E[XY] = E[X]E[Y]$ is not in general true, more will be said about it in future lectures. You will see the applications of linearity of expectation throughout this course, below we look at examples.

A simple example: Remember the n envelope and letter problem. We have n letters and n envelopes, numbered from 1 to n. If we assign the letters to envelopes randomly, how many of the letters do we expect to get into the correct envelope?

Let C be the random variable which captures the number of correct letters to envelopes. The expected value of C is

$$E[C] = \sum_{0 \le k \le n} k P(C = k).$$

We have calculated $P(C = k)$ before (inclusion-exclusion), it is a mess and it will be a further mess to put it in the summation. Linearity of expectation gives us a way out. Suppose C_i denotes the random variable which is 1 when i-th letter goes to the correct envelope and 0 otherwise.

Note 4. Such random variables are called indicator random variables which take value 0 or 1. They *indicate* if a certain event happens or not.

Exercise 12. What is C in terms of C_i's?

Some thought shows that C can be expressed as summation of C_i's. The expectation of C can be computed easily now.

$$\begin{aligned} E[C] =& E[C_1 + C_2 + \cdots + C_n] \\ &= E[C_1] + E[C_2] + \cdots + E[C_n] \\ &= nE[C_1] \\ &= 1 \end{aligned} \tag{3}$$

For the last two equations, you need the following exercise.

Exercise 13. Show that $E[C_i] = P[i\text{-th letter goes in correct envelope}] = 1/n$.

It is not just that the calculation is simple, the same calculation goes through with much lesser assumptions on letter to envelope problem. We just need that each letter goes to correct envelope with probability $1/n$, it doesn't matter if all permutations are considered or equally likely.

Another example: Suppose you want to collect stickers which accompany your favorite chewing-gum. There are n different stickers numbered 1 to n. Every time you buy a chewing-gum, one sticker comes out of n with equal probability. What is the expected number of chewing-gums you need to buy to collect all possible stickers.

Let T be the random variable which counts the number of packets to be bought to collect all the different n stickers. We will define T as sum of random variables, these random variables need to be defined carefully.

Let S_1 be the random variable that we get first distinct sticker (clearly $S_1 = 1$). Let S_2 be the extra number of chewing-gums for getting second different sticker, similarly define S_k. Intuitively, you might want to define S_1 as the number of days to collect sticker number 1 and so on. You will show in the assignment that this strategy will not work. So, S_i is defined as the number of days to get i-th different sticker (sticker number could be anything) after we got $i-1$ different sticker.

We need to calculate $E[T] = E[S_1 + S_2 + \cdots + S_n]$. By linearity of expectation, we only need to worry about $E[S_k]$. The probability that $S_k = r$ is,

$$Pr(S_k = r) = ((k-1)/n)^{r-1}\left(1 - \frac{k-1}{n}\right).$$

Exercise 14. Show that $E[S_k] = \frac{n}{n-(k-1)}$.

This implies that the expected number of days needed to collect all stickers is $E[T] = \sum_k \frac{n}{n-(k-1)}$. Since we know that $1 + 1/2 + 1/3 + \cdots$ approach $\ln n$, $E[T]$ is around $n \ln n$.

2.2 Conditional expectation

The scope of a random variable X could be restricted to an event B, denoted by $X|_B$. This means that we can define *conditional distribution* of $X|_B$.

Exercise 15. What should be $P_{X|_B}(x)$?

The conditional probability distribution $P(X = x|B)$ should give the probability of $X = x$ given that event B has happened. Intuitively,

$$P_{X|_B}(x) = P(X = x|B) = \frac{P(B \cap X^{-1}(x))}{P(B)}.$$

Similarly, we can also talk about conditional expectation, similar to the expectation of a random variable.

Exercise 16. What should it mean? How should you define it?

It seems that the conditional expectation, $E[X|A]$, should be the *average* value of random variable X given that event A has happened. Our first guess would be, $E[X|A] = \sum_{\omega \in A} P(\omega)X(\omega)$. The only thing to notice, it should be $P(\omega|A)$ and not just $P(\omega)$. This gives us the correct formula,

$$E[X|A] = \sum_{\omega \in A} P(\omega|A)X(\omega) = \frac{1}{P(A)} \sum_{\omega \in A} P(\omega)X(\omega) = \sum_{x \in Range(X)} xP(X = x|B).$$

Exercise 17. Why did we get the second equality?

We can also generalize the partition formula, given disjoint events $B_1, B_2, \cdots, B_k$,

$$E[X] = \sum_{i=1}^{k} P(B_i)E[X|B_i].$$

Let us look at a problem. Suppose a miner, stuck in a mine, has to choose one door out of three. One door takes him back to where he is but it will waste his 7 hours. Second door takes him out in 3 hours. The last one again get him back to the starting point, but requires 5 hours. If he chooses a door uniformly, what is the expected time in which he will get out of the mine?

Let time taken to get out be the random variable X. He chooses door D_1, D_2, D_3 (use the same names for corresponding events) with equal probability. Then,

$$E[X] = \sum_{i=1}^{3} P(D_i)E[X|D_i].$$

We know that $P(D_i) = 1/3$ and $E[X|D_2] = 3$.

Exercise 18. What is $E[X|D_1]$ and $E[X|D_3]$?

Substituting the values in the equation.

$$3E[X] = 7 + E[X] + 3 + 5 + E[X].$$

We expect that miner will get out in 15 hours.

The concept of conditional probability allows us to have an expression for joint probability distribution function too. Remember that the joint distribution for two random variables X, Y was defined as $p_{X,Y}(x, y) = P(X = x, Y = y)$. The right hand side is the probability that both events, $X = x$ and $Y = y$, happen. Using conditional probability,

$$P(X = x, Y = y) = P(X = x|Y = y)P(Y = y).$$

Exercise 19. Extend this expression to calculate $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$.

It seems like the joint distribution is easy to calculate if X, Y are independent. What should it mean for two random variables to be independent?

Remember, a random variable X is a mapping from sample space to real numbers, $X : \Omega \to \mathbb{R}$. Intuitively, all events of kind $X = x$ and $Y = y$ should become independent.

Formally, two random variables X and Y are independent iff,

$$P(X = x \cap Y = y) = P(X = x)P(Y = y) \quad \forall x, y.$$

Here x, y are in the range of X, Y respectively.

When X, Y are independent, $P(X = X|Y = y) = P(X = x)$ and we get $P(X = x, Y = y) = P(X = x)P(Y = y)$. Otherwise, we need the joint distribution explicitly or the relation between X and Y.

2.3 Variance

We have already seen one aggregate measure of a random variable, expectation. Clearly, expectation captures only a very small amount of information about the random variable. It was emphasized before, expectation does not imply that we get $E[X]$ with high probability. In other words, a random variable which always takes value 0 has same expectation as the one which takes value -2 and 2 with equal probability (even one which takes values 1000 and -1000 with same probability). Expectation does not distinguish between these cases.

This gives rise to another measure of interest, called the *variance* of a random variable. The idea is to measure how far X could be from $E[X]$.

Your first guess might be $E[X - E[X]]$, but we need to take care of the signs of random variable $X - E[X]$. The *variance* of a random variable X is defined as,

$$Var[X] := E[(X - E[X])^2).$$

Exercise 20. Show that

$$Var[X] = E[X^2] - (E[X])^2.$$

You can easily calculate the variance of a random variable taking value $\alpha, -\alpha$ with equal probability. It increases rapidly as the value of α increases, even though the expectation remains same.

A very related quantity is called the *standard deviation* and is defined as the square root of the variance.

$$\sigma(X) := \sqrt{Var[X]} = \sqrt{E[X^2] - E[X]^2}.$$

Define a random variable X to be 0 for tails and 1 for head, where coin gets head with probability p. This is called a Bernoulli random variable with parameter p.

Exercise 21. What is the expectation of this random variable? What is the expectation of X^2?

From this exercise, it is easy to find the standard deviation of X,

$$\sigma(X) = \sqrt{p - p^2} = \sqrt{p(1-p)}.$$

Exercise 22. For what value of p is this maximum? Does it agree with your intuition?

3 Some important random variables and their distribution

Some of the well-known random variables and their distributions occur quite frequently in practice (many a times in disguised form). We look at them below and to get more experience, also calculate their expectation.

Bernoulli random variable This is one of the simplest random variables, you have already seen it. Think of a biased coin which lands head with probability p (sometimes called the success probability). Bernoulli random variable, X, is obtained by assigning 1 to heads and 0 to tails.

Exercise 23. What is the sample space Ω? Write the random variable and its probability mass functions explicitly (as a function).

Now, the expectation can be calculated,

$$E[X] = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = p.$$

We calculated its variance in the previous section.

Binomial random variable The coin from the previous example can be tossed multiple times. Define X to be the random variable which counts the number of heads when the coin is tossed n times. Again, we can write the probability mass function,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Note 5. This expression works for k even when it is larger than n.

Exercise 24. Calculate the expectation value of X?

You probably applied the formula and found the expectation to be np by applying your knowledge of binomial coefficients. How about a totally different way? We can express the binomial random variable X as the sum of n Bernoulli random variables $X_1, X_2, \cdots, X_n$. So,

$$E[X] = E[X_1 + X_2 + \cdots + X_n] = \sum_{i=1}^{n} E[X_i] = np.$$

We see that linearity of expectation makes it easier to calculate the expectation. The variance could also be calculated similarly by realizing that $X_1, X_2, \cdots$ are not related to each other and are *independent.* We will see a formal definition later and the property of variance when the terms are independent. For now, you will calculate the variance directly (assignment question).

Geometric random variable Again, we have a biased coin with probability p of getting head. Suppose, we toss the coin till we get a head. Geometric random variable, X, counts the number of trials for such an experiment.

Exercise 25. What is the sample space and probability distribution function for such an experiment?

With some thought,

$$P(X = k) = (1-p)^{k-1}p.$$

This allows us to calculate the expectation,

$$E[X] = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = 1/p.$$

Exercise 26. How do you derive the last equality?

You can calculate the variance of geometric random variable by computing $E[X(X-1)]$, it comes out to be $\frac{1-p}{p^2}$.

We can also stop the experiment only after getting n heads. The random variable which counts the number of trials in this case, is called a negative binomial random variable.

Exercise 27. Calculate the expected value of a negative binomial random variable.

3.1 Continuous random variables

We can also define a continuous random variable, when the sample space is uncountable. For instance, $\Omega = \mathbb{R}$ could be the sample space and then the random variable will be a function $X : \Omega \to \mathbb{R}$. What would be the probability mass function? We discussed that $P(X = r)$ for any real r should be zero.

In case of continuous random variables, we define probabilities of intervals. Specifically, the events of interest are $X \leq r := \{\omega \in \Omega : X(\omega) \leq r\}$. For this definition to make sense, the event $X \leq r$ should be in sigma field. This comes from conditions imposed through *measure theory*, and that discussion is beyond the scope of this course. For us, we will assume that the events $X \leq r$ are part of our sigma field.

Now we can define the probability distribution function to be $F(x) := P(X \leq x)$ in general. Throughout this course we assume that F is going to be continuous and there exist a non-negative function $f : \mathbb{R} \to \mathbb{R}$, such that,

$$F(a) = \int_{-\infty}^{a} f(x)\ dx \textit{ and specifically } \int_{-\infty}^{\infty} f(x)\ dx = 1.$$

The function f is called the *probability density function (pdf)* of X and is the analog of probability mass function of a discrete random variable. Though, it is not correct to say that $f(x)$ is the probability of getting

x, we know it to be zero (instead, intuitively we can think of it as the probability of the interval $(x, x+dx)$). The integration of the pdf of X gives us the probability in an interval.

$$P(a \leq x \leq b) = \int_a^b f(x)\ dx.$$

The expectation of X can be defined by,

$$E[X] = \int_{-\infty}^{\infty} x f(x)\ dx.$$

Like before, variance will be defined similar to the discrete case, $Var[X] = E[(X - E[X])^2]$. Let us see some examples.

Exponential random variable An exponential random variable with parameter λ is defined by the pdf,

$$f(x) = \lambda e^{-\lambda x} \quad for\ x \geq 0.$$

We can calculate the expectation,

$$E[X] = \int_0^{\infty} x \lambda e^{-\lambda x}\ dx = 1/\lambda.$$

By calculating $E[X^2]$, we get that $Var[X] = 1/\lambda^2$.

Normal/Gaussian random variable Consider a Bernoulli random variable with success probability p. We call a sample to be a draw of n values from this distribution. In other words, our sample size is n. If we get $X_1, X_2, \cdots, X_n$ as draws, say that $(X_1 + X_2 + \cdots + X_n)/n$ is the value of the sample.

Exercise 28. If we draw n such samples, how will the value of these samples be distributed?

This is nothing but the binomial random variable divided by n (for n tosses), we know that the expectation is p. Let us plot this distribution for increasing values of n.

You see that the shape looks more and more like a bell. It can be formally shown that the distribution (after some shifting and scaling) approaches the continuous random variable,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Such a distribution is called a *standard normal or Gaussian distribution.* Surprisingly, you can start with any distribution of your choice. If you look at the distribution of mean of sample size n, the distribution looks like a normal random variable as n tends to infinity. This is known as *central limit theorem*, I strongly urge you to read more about it.

Another way to state central limit theorem is: we might not know a distribution coming from nature, still if we look at the aggregate behavior (mean of the draws), it looks like a normal distribution. Hence, this distribution is hugely popular in statistics.

Exercise 29. What is the expected value of normal random variable? Can you see the role of symmetry?

By a more difficult integration, you can show that $E[X^2] = 1$, where X is standard normal random variable. This shows that the variance (and standard deviation) for a standard normal random variable is 1.

We can shift and scale the Gaussian random variable (change the expectation and standard deviation), the Gaussian distribution with expectation μ and standard deviation σ is defined by the pdf,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

Bayesian parameter estimation This example is taken from *Notes on Bayesian Learning* by Padhraic Smyth (University of California, Irvine) [1].

The problem of parameter estimation is: given that data is supposed to follow a known distribution with parameter θ (for instance, mean and variance for normal distribution), if we observe some data, how should our *belief* change about θ. There can be multiple ways to concretely formulate this problem. In case of Bayesian learning, we assume that θ itself has some known distribution (it is a random variable). Notice that the distribution of θ need not be same as the distribution for the data (the prior distribution of the parameters, before observing data, is completely known). We need to find the new belief (distribution of θ) after observing the data.

To set up the notation, we will be given the prior distribution of parameter θ, say $p(\theta)$. Then we observe some data D, and our task is to find new distribution $p(\theta|D)$. You can guess that it is a standard application of Bayes' theorem. It will be,

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.$$

$p(\theta)$ is called the prior distribution, and $p(\theta|D)$ is called the posterior distribution. Our task is to update prior to get posterior, given the data. The normalization factor $p(D)$ can be calculated (most of the times) by observing that $p(\theta|D)$ is a probability distribution.

Let us take an example. Suppose our data $D \in \{0,1\}^n$ is coming from Bernoulli distribution with parameter θ, i.e., the probability of success is θ. The data can be interpreted as a string of success and failures. How should we find the distribution of θ.

Exercise 30. What should be the prior $p(\theta)$?

One of the natural choice is the uniform distribution over $[0,1]$. We will take a slightly more general form, the motivation will become clear later. We assume that θ is distributed as a Beta distribution,

$$p(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Here α,β are the parameters of Beta distribution and are known before we observe the data; $B(\alpha,\beta)$ is the normalization constant.

Exercise 31. If we initially believed that θ was uniformly distributed in $[0,1]$, what should be α,β?

Now, it is not difficult to find posterior distribution, but first notice that

$$p(D|\theta) = \theta^s(1-\theta)^{n-s},$$

where s is the number of successes in D.

Now,

$$p(\theta|D) \propto \theta^s(1-\theta)^{n-s}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Exercise 32. Show that the posterior also has a Beta distribution with parameters $\alpha+s, \beta+n-s$.

4 Buffon's needle problem

Let us consider another problem in a continuous domain. Suppose you are given a board with lines drawn horizontally at a distance of 1 unit length, parallel to each other. If we drop a needle of unit length on this board, what is the expected value of intersections of the needle with lines on the board?

Again, what is the sample space, random variable and its probability density function? To not worry about the boundary conditions, assume that the board is *infinite*. Clearly the random variable is the number

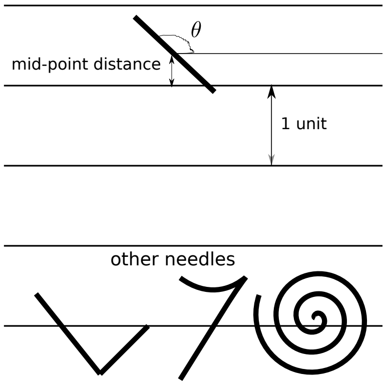

Fig. 1. Dropping a needle on an infinite board

of intersections. We can think of the dropping the needle uniformly at some distance from a line (mid-point of the needle between 0 and $1/2$) with some angle uniformly (between $\pi/2$ and $-\pi/2$).

In other words, we choose a point uniformly in the rectangle of area $1/2 \times \pi = \pi/2$. This defines our sample space. What is the pdf? Since we are dropping the needle uniformly, define the pdf to be,

$$f(y,\theta) = 2/\pi \quad 0 \leq y \leq 1/2, -\pi/2 \leq \theta \leq \pi/2.$$

Exercise 33. Evaluate the double integral $\int_0^{1/2} \int_{pi/2}^{\pi/2} f(y,\theta)\ dy\ d\theta$ and show that it is a valid pdf.

Notice that the needle intersects one line if $y \leq 1/2 \cos\theta$. We can ignore the case when midpoint is at a distance of $1/2$ (why?).

Calculating the expectation,

$$E[\textit{number of intersections}] = \int_{-\pi/2}^{\pi/2} \int_0^{1/2\cos\theta} f(y,\theta)\ dy\ d\theta.$$

This is an elementary integral and the value comes out to be $2/\pi$. We can make the problem much more difficult. What if the needle is not of unit length? How about the case when needle is circular or bent or of shape W?

Seems like it is going to be a very difficult integration, even if we can express the expectation as an integral. Here, we will see the power of linearity of expectation.

Divide the needle into very small parts. Clearly the total number of intersections (say random variable Y) can be expressed as the sum of intersections of each individual part (say random variable X_i). By linearity of expectation,

$$E[Y] = \sum_i E[X_i].$$

If we divide the needle into very very small parts, shows that the expectation is proportional to the length of the needle.

$$E[\textit{number of intersections}] = cL \quad (\textit{L is the length of the needle}).$$

It might take some time to convince yourself of the above statement, take your time. Linearity of expectation seemed natural and easy. Still, if someone said that the expected number of intersections of the needle does not depend on its shape, it would not sound believable. Remember that the needle could be circular, bent or of any given shape.

We have one more job to finish. How to find this absolute constant c? If we know the number of intersections of any needle and its length, it will give us the constant c.

Exercise 34. Can you think of a case where you know the expectation and length of the needle?

We already know one case. The expectation of a straight needle of length 1 is $2/\pi$ giving $c = 2/\pi$. There is an easier case, we don't even need to evaluate the first integral. Take a circular needle of diameter 1.

Exercise 35. What is the expected number of intersections in this case?

This gives us the final formula,

$$E[\textit{number of intersections}] = \frac{2}{\pi}L \quad (\textit{L is the length of the needle}).$$

5 Assignment

Exercise 36. In a group of 23 people, we ask birthday of everyone. Define the random variable X to be the number of pairs whose birthdays match. What is $Pr(X \geq 1)$?

Exercise 37. Suppose you pick two cards from a deck with cards numbered from 1 to 1000. What is the expected value of the greater number?

Exercise 38. Suppose you roll a dice twice and X be the number on the first roll and Y be the number on the second roll. Assuming $P(X = x, Y = y) = P(X = x)P(y = y)$, what is the k for which $P(X + Y = k)$ is maximized?

Exercise 39. Let X be a random variable with $Pr(X = 1) = p$ and $Pr(X = -1) = 1 - p$. Find $E[X^n]$.

Exercise 40. For the sticker collection problem in the linearity of expectation, let us look at a different solution. Say T_i be the random variable which counts the packets needed to collect i^{th} sticker. Then $E[T_i] = n$, and $E[T] = \sum_i E[T_i] = n^2$. Is this argument correct, if not, what is wrong with this argument?

Exercise 41. Find the variance of binomial distribution directly by finding out $E[X(X - 1)]$.

References

1. P. Smyth. Notes on bayesian learning, 2019.