Probability

Onkar Pandit

24 February 2022

Last updated on June 7, 2022.

Contents

1 Introduction

When constructing automated reasoning systems, classical Boolean logic does not allow us to express certain forms of plausible reasoning. Consider the following scenario: We observe that A is false. We find B becomes less plausible, although no conclusion can be drawn from classical logic. We observe that B is true. It seems A becomes more plausible. We use this form of reasoning daily. We are waiting for a friend, and consider three possibilities: H_1, she is on time; H_2, she has been delayed by traffic; and H_3, she has been abducted by aliens. When we observe our friend is late, we must logically rule out H_1. We also tend to consider H_2 to be more likely, though we are not logically required to do so. Finally, we may consider H_3 to be possible, but we continue to consider it quite unlikely. How do we conclude H_2 is the most plausible answer? Seen in this way, probability theory can be considered a generalization of Boolean logic. Logic provides a set of formal rules for determining what propositions are implied to be true or false given the assumption that some other set of propositions is true or false. Probability theory provides a set of formal rules for determining the likelihood of a proposition being true given the likelihood of other propositions.

We don't really answer the question 'What is probability?' Nobody has a really good answer to this question. Some people think of it as 'limiting frequency'. That is, to say that the probability of getting heads when a coin is tossed means that, if the coin is tossed many times, it is likely to come down heads about half the time. But if you toss a coin 1000 times, you are not likely to get exactly 500 heads. You wouldn't be surprised to get only 495. But what about 450, or 100?

Some people would say that you can work out probability by physical arguments, like the one we used for a fair coin. But this argument doesn't work in all cases, and it doesn't explain what probability means.

Some people say it is subjective. You say that the probability of heads in a coin toss is $1/2$ because you have no reason for thinking either heads or tails more likely; you might change your view if you knew that the owner of the coin was a magician or a con man. But we can't build a theory on something subjective.

We regard probability as a mathematical construction satisfying some axioms (devised by the Russian mathematician A. N. Kolmogorov). We develop ways of doing calculations with probability, so that (for example) we can calculate how unlikely it is to get 480 or fewer heads in 1000 tosses of a fair coin.

In machine learning and statistics, there are two major interpretations of probability: the Bayesian and frequentist interpretations. The Bayesian interpretation uses probability to specify the degree of uncertainty that the user has about an event. It is sometimes referred to as "subjective probability" or "degree of belief". The frequentist interpretation considers the relative frequencies of events of interest to the total number of events that occurred. The probability of an event is defined as the relative frequency of the event in the limit when one has infinite data.

Nearly all activities require some ability to reason in the presence of uncer-

tainty. In fact, beyond mathematical statements that are true by definition, it is difficult to think of any proposition that is absolutely true or any event that is absolutely guaranteed to occur.

There are three possible sources of uncertainty:

1. Inherent stochasticity in the system being modelled. For example, most interpretations of quantum mechanics describe the dynamics of subatomic particles as being probabilistic. We can also create theoretical scenarios that we postulate to have random dynamics, such as a hypothetical card game where we assume that the cards are truly shuffled into a random order.

2. Incomplete observability. Even deterministic systems can appear stochastic when we cannot observe all of the variables that drive the behaviour of the system. For example, in the Monty Hall problem, a game show contestant is asked to choose between three doors and wins a prize held behind the chosen door. Two doors lead to a goat while a third leads to a car. The outcome given the contestant's choice is deterministic, but from the contestant's point of view, the outcome is uncertain.

3. Incomplete modelling. When we use a model that must discard some of the information we have observed, the discarded information results in uncertainty in the model's predictions. For example, suppose we build a robot that can exactly observe the location of every object around it. If the robot discretizes space when predicting the future location of these objects, then the discretization makes the robot immediately become uncertain about the precise position of objects: each object could be anywhere within the discrete cell that it was observed to occupy.

In many cases, it is more practical to use a simple but uncertain rule rather than a complex but certain one, even if the true rule is deterministic and our modelling system has the fidelity to accommodate a complex rule. For example, the simple rule "Most birds fly" is cheap to develop and is broadly useful, while a rule of the form, "Birds fly, except for very young birds that have not yet learned to fly, sick or injured birds that have lost the ability to fly, flightless species of birds including the cassowary, ostrich and kiwi $\cdots$" is expensive to develop, maintain and communicate, and after all of this effort is still very brittle and prone to failure.

2 Basics

Sample space The general setting is: We perform an experiment which can have a number of different outcomes. The sample space is the set of all possible outcomes of the experiment. We usually call it $\mathcal{S}$.
For example, sample space for tossing coin three times is given as:

$$\mathcal{S} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Events An event is a subset of $\mathcal{S}$. We can specify an event by listing all the outcomes that make it up. In the above example, let A be the event 'more heads than tails' and B the event 'heads on last throw'. Then,

$$A = \{HHH, HHT, HTH, THH\}$$
$$B = \{HHH, HTH, THH, TTH\}$$

2.1 Axioms

A function P that assigns a real number $P(A)$ to each event A is a probability distribution or a probability measure if it satisfies the following three axioms:

1. For any event A, we have $P(A) \geq 0$.
2. For sample space $\mathcal{S}$, $P(\mathcal{S}) = 1$.
3. A number of events, say $A_1, A_2, \cdots, A_n$ are called mutually disjoint or pairwise disjoint if $A_i \cap A_j = \emptyset$ for any two of the events A_i and A_j; that is, no two of the events overlap. Then,

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

There are many interpretations of $P(A)$. The two common interpretations are frequencies and degrees of beliefs. In the frequency interpretation, $P(A)$ is the long run proportion of times that A is true in repetitions. For example, if we say that the probability of heads is $1/2$, we mean that if we flip the coin many times then the proportion of times we get heads tends to $1/2$ as the number of tosses increases. An infinitely long, unpredictable sequence of tosses whose limiting proportion tends to a constant is an idealization, much like the idea of a straight line in geometry. The degree-of-belief interpretation is that $P(A)$ measures an observer's strength of belief that A is true. In either interpretation, we require that Axioms 1 to 3 hold. The difference in interpretation leads to two schools of inference: the *frequentist* and the *Bayesian* schools.

Few other properties which can be derived from the basic probability axioms which are as follows:

1. $P(A') = 1 - P(A)$
2. Inclusion-exclusion principle: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This can be generalized to more events as:

$$P(A_1 \cup A_2 \cup \cdots A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n} \sum_{j=i+1}^{n} P(A_i \cap A_j) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} P(A_i \cap A_j \cap A_k) - \cdots$$

3. Two events A and B are said to be independent if $P(A \cap B) = P(A)P(B)$. More generally, let $A_1, A_2, \cdots, A_n$ be mutually independent. Then

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = \Pi_{i=1}^{n} P(A_i)$$

.

2.2 Conditional Probability

Let us see, conditional dependence of one vent on another event with an example.

Alice and Bob are going out to dinner. They toss a fair coin 'best of three' to decide who pays: if there are more heads than tails in the three tosses then Alice pays, otherwise Bob pays. Clearly each has a 50% chance of paying. The sample space is

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

and the events 'Alice pays' and 'Bob pays' are respectively

$$A = \{HHH, HHT, HTH, THH\},$$
$$B = \{HTT, THT, TTH, TTT\}$$

They toss the coin once and the result is heads; call this event E. How should we now reassess their chances? We have

$$E = \{HHH, HHT, HTH, HTT\}$$

, and if we are given the information that the result of the first toss is heads, then E now becomes the sample space of the experiment, since the outcomes not in E are no longer possible. In the new experiment, the outcomes 'Alice pays' and 'Bob pays' are

$$A \cap E = \{HHH, HHT, HTH\},$$
$$B \cap E = \{HTT\}$$

From this we can see that

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

A generic formula for probability of ' event A occurring given the event B' is given as

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

Suppose there is a certain event which depends on more than one event, then the generic form of conditional probability is given as:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_n|A_1, A_2, \cdots, A_{n-1}) \cdots P(A_2|A_1)P(A_1) \tag{2}$$

2.3 Theorem of Total Probability

Let $A_1, A_2, \cdots, A_n$ form a partition of the sample space with $P(A_i) \neq 0 \ \ \forall i$, and let B be any event. Then the theorem states that:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \tag{3}$$

Another way to write it: $P(B) = P(B|A)P(A) + P(B|A')P(A')$.
From Eq. 3 we can also write:

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) \tag{4}$$

2.4 Bayes Theorem

Let A_j and B be events with non-zero probability. Then

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} \tag{5}$$

In general, $P(A_j)$ is called the prior probability of A_j and $P(A_j|B)$ the posterior probability of A_j. This is because first we assume the probability distribution as $P(A_j)$ and then update our beliefs, posterior.

In general, following form of Bayes' is used more often:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')} \tag{6}$$

3 Random variables

A random variable is neither random nor a variable. It is a function on a sample space which produces numbers such as random variable $X : \mathcal{S} \to R$. For example, selection of a student from the class at random and measure his or her height in centimetres. Here, the sample space is the set of students; the random variable is *height*, which is a function from the set of students to the real numbers: h(S) is the height of student S in centimetres.

The probability rules which we defined for an event can be extended for random variables as when these rules are applied, they are eventually applied for a particular value of the random variable, which is nothing but an event. For example, the Bayes' theorem can be extended to random variables X, Y as

$$P(X|Y) = \frac{P(Y|X)P(X)}{\sum_X P(X, Y)} \tag{7}$$

In the above equation, the X, Y are kind of placeholders, which are going to take concrete values, but generalized relation between these random variables

hold. Suppose, $X = \{a, b, c\}$ and $Y = \{1, 2\}$ then the bayes equation will be defined over six different combinations: $\mathcal{S} = \{(a, 1), (a, 2), (b, 1), (b, 2), (c, 1), (c, 2), \}$. For a particular pair $(c, 2)$ the value is:

$$P(X = c|Y = 2) = \frac{P(Y = 2|X = c)P(X = c)}{\sum_{(i,j)\in\mathcal{S}} P(Y = j|X = i)P(X = i)} \tag{8}$$

For random variables X, Y:

1. *Sum Rule*: $P(X) = \sum_Y P(X, Y)$
2. *Product Rule*: $P(X, Y) = P(X|Y)P(Y)$

We will be explicit about random variables and the values they take. Hence, recall that we use capital letters X, Y to denote random variables and small letters x, y to denote the values in the target space $\mathcal{T}$ that the random variables take. We will explicitly write pmfs of discrete random variables X as $P(X = x)$. For continuous random variables X, the pdf is written as $f(x)$ and the cdf is written as $F_X(x)$.

Expectation Expected value of random variable X is given as:

$$E(X) = \sum_{i=1}^{n} x_i p(x_i) \tag{9}$$

where x_i are different values r.v. X can take and $p(x_i)$ is the associated probability for it.

Further, expectation can be defined over $f(x)$ where $x \in X$. It's value is calculated as:

$$E(f(x)) = \sum_{i=1}^{n} f(x_i)p(x_i) \tag{10}$$

We can also define expectations with condition on some other random variable: *conditional expectations* as,

$$E(f(x|y = y)) = \sum_{i=1}^{n} f(x_i)p(x_i|y) \tag{11}$$

Variance Variance of random variable X is given as:

$$Var(X) = E((X - \mu)^2) \tag{12}$$

$$Var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i) \tag{13}$$

$$Var(X) = E(X^2) - (E(X))^2 \tag{14}$$

Standard deviation is calculated by taking square root of variance, denoted as σ.

3.1 Discrete random variables

Discrete random variables can take on either a finite or at most a countably infinite set of discrete values (for example, the integers). Their probability distribution is given by a probability mass function which directly maps each value of the random variable to a probability. For example, the value of x_1 takes on the probability p_1, the value of x_2 takes on the probability p_2, and so on. The probabilities p_i must satisfy two requirements: every probability $0 \geq p_i \leq 1$, and the sum of all the probabilities is 1.

Probability mass function The *probability mass function* of a discrete random variable X is the function, formula or table which gives the value of $P(X = a)$ for each element a in the target set of X. If X takes only a few values, it is convenient to list it in a table; otherwise we should give a formula if possible. The standard abbreviation for 'probability mass function' is p.m.f[1].

For example, suppose a fair coin is tossed three times. The random variable X gives the number of heads recorded. The possible values of X are 0, 1, 2, 3, and its p.m.f. is Table ??:

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Joint Probability Mass Function If two random variables X and Y are not independent, then knowing the p.m.f. of each variable does not tell the whole story. The *joint probability mass function* (or joint p.m.f.) of X and Y is the function which produces probability for each a_i of X and each value b_j of Y, which is denoted as $P(X = a_i, Y = b_j)$. These values if are manageable then can be presented in a table. We arrange the table so that the rows correspond to the values of X and the columns to the values of Y . Note that summing the entries in the row corresponding to the value a_i gives the probability that $X = a_i$; that is, the row sums form the p.m.f. of X. Similarly the column sums form the p.m.f. of Y. The row and column sums are sometimes called the *marginal distributions* or *marginals*.

Cumulative Distribution Function Let X be a random variable taking values $a_1, a_2, \cdots, a_n$. We assume that these are arranged in ascending order: $a_1 < a_2 < \cdots < a_n$. The cumulative distribution function, or c.d.f., of X is given by

$$F_X(a_i) = P(X \leq a_i) \tag{15}$$

[1]There is a fairly common convention in probability and statistics that random variables are denoted by capital letters and their values by lower-case letters. In fact, it is quite common to use the same letter in lower case for a value of the random variable; thus, we would write $P(X = x)$ in the above example. But remember that this is only a convention, and you are not bound to it.

We see that it can be expressed in terms of the p.m.f. of X as follows:

$$F_X(a_i) = P(X = a_1) + \cdots + P(X = a_i) = \sum_{j=1}^{i} P(X = a_j) \tag{16}$$

We can also write, pmf with cdf as:

$$P(X = a_i) = F_X(a_i) - F_X(a_{i-1}) \tag{17}$$

Independence of random variable Let X be a random variable taking the values $a_1, \cdots, a_n$, and let Y be a random variable taking the values $b_1, \cdots, b_m$. We say that X and Y are independent if, for any possible values i and j, we have

$$P(X = a_i, Y = b_j) = P(X = a_i)P(Y = b_j)$$

From these definitions, we can state following facts.

1. $E(X + Y) = E(X) + E(Y)$
2. If X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$.
3. E(X + c) = E(X) + E(C) = E(X) + c
4. Var(X + c) = Var(X) + Var(C) = Var(X)
5. E(cX) = c E(X)
6. Var(cX) = c^2Var(X)

We prove the first theorem, the second one can be proved in a similar fashion. Prove $E(X + Y) = E(X) + E(Y)$.

$$\begin{aligned}
E(X+Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m}(a_i + b_j)P(X = a_i, Y = b_j) \\
&= \sum_{i=1}^{n} a_i \sum_{j=1}^{m} P(X = a_i, Y = b_j) + \sum_{j=1}^{m} b_j \sum_{i=1}^{n} P(X = a_i, Y = b_j) \\
&= \sum_{i=1}^{n} a_i P(X = a_i) + \sum_{j=1}^{m} b_j P(Y = b_j) \\
E(X+Y) &= E(X) + E(Y)
\end{aligned}$$

3.2 Popular discrete probability distributions

We now look at five types of discrete random variables, each depending on one or more parameters. We describe for each type the situations in which it arises, and give the p.m.f., the expected value, and the variance.

3.2.1 Bernoulli Distribution

A Bernoulli random variable is the simplest type of all. It only takes two values, 0 and 1. So its p.m.f. is $P(X = 1) = p$ and $P(X = 0) = 1 - p$. For a Bernoulli random variable X, we sometimes describe the experiment as a 'trial', the event $X = 1$ as 'success', and the event $X = 0$ as 'failure'. For example, if a biased coin has probability p of coming down heads, then the number of heads that we get when we toss the coin once is a $Bernoulli(p)$ random variable.

Calculation of the expected value and variance of a Bernoulli random variable is easy. Let $X \sim Bernoulli(p)$ (which denotes random variable X follows $Bernoulli(p)$ distribution). Then,

$$E(X) = 0(1 - p) + 1p = p \tag{18}$$

$$Var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1 - p) \tag{19}$$

3.2.2 Binomial Distribution

Remember that for a Bernoulli random variable, we describe the event $X = 1$ as a 'success'. Now a binomial random variable counts the number of successes in n *independent* trials each associated with a $Bernoulli(p)$ random variable.

For example, suppose that we have a biased coin for which the probability of heads is p. We toss the coin n times and count the number of heads obtained. This number is a $Bin(n, p)$ random variable. A $Bin(n, p)$ random variable X takes the values $0, 1, 2, \cdots, n$, and the p.m.f. of X is given by

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{20}$$

where $0 \leq k \leq n$.
Note that the Bernoulli distribution is just the binomial distribution with n = 1. If $X \sim Bin(n, p)$, then

$$E(X) = np, \qquad Var(X) = np(1 - p) \tag{21}$$

Let us see, how it can be proved. Suppose, we have a coin with probability p of coming down heads, and we toss it n times and count the number X of heads. Then $X \sim Bin(n, p)$ random variable. Let $X_k \sim Bernoulli(p)$ be the random variable which corresponds to each n toss. Then we can write

$$X = X_1 + X_2 + \cdots + X_n$$

as it will give the number of heads. Then expectation of X can be written as: $E(X) = E(X_1) + E(X_2) + \cdots + E(X_n)$ which is nothing but np. Similarly, variance can be written as $Var(X) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$ as all the trials are independent. Thus, $Var(X) = np(1 - p)$.

3.2.3 Hypergeometric Distribution

Suppose that we have N balls in a box, of which M are red. We sample n balls from the box without replacement. Let the random variable X be the number of red balls in the sample. Such an X is called a *hypergeometric* random variable $Hg(n, M, N)$. The random variable X can take any of the values $0, 1, 2, \cdots, n$. Its p.m.f. is given by the formula

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \tag{22}$$

For the number of samples of n balls from N is $\binom{N}{n}$; the number of ways of choosing k of the M red balls and $n-k$ of the $N-M$ others is $\binom{M}{k}.\binom{N-M}{n-k}$; and all choices are equally likely.

The expected value and variance of a hypergeometric random variable are as follows:

$$E(X) = n\frac{M}{N} \qquad Var(X) = n\frac{M}{N}\frac{N-M}{N}\frac{N-n}{N-1} \tag{23}$$

You should compare these to the values for a binomial random variable. If we let $p = \frac{M}{N}$ be the proportion of red balls in the hat, then $E(X) = np$, and $Var(X) = npq$ multiplied by a 'correction factor' $\frac{N-n}{N-1}$. In particular, if the numbers M and $N-M$ of red and non-red balls in the hat are both very large compared to the size n of the sample, then the difference between sampling with and without replacement is very small, and indeed the 'correction factor' is close to 1. So we can say that $Hg(n, M, N)$ is approximately $Bin(n, \frac{M}{N})$ if n is small compared to M and $N-M$.

3.2.4 Geometric Distribution

The geometric random variable is like the binomial but with a different stopping rule. We have again a coin whose probability of heads is p. Now, instead of tossing it a fixed number of times and counting the heads, we toss it until it comes down heads for the first time, and count the number of times we have tossed the coin. Thus, the values of the variable are the positive integers $1, 2, 3, \cdots$ (In theory we might never get a head and toss the coin infinitely often, but if $p > 0$ this possibility is 'infinitely unlikely', i.e. has probability zero, as we will see.) We always assume that $0 < p < 1$. More generally, the number of independent Bernoulli trials required until the first success is obtained is a geometric random variable.

The p.m.f of a Geom(p) random variable is given by

$$P(X = k) = (1-p)^k p \tag{24}$$

Expected value and variance is given as:

$$E(X) = \frac{1}{p} \qquad Var(X) = \frac{1-p}{p^2} \tag{25}$$

3.2.5 Poisson Distribution

The Poisson random variable, unlike the ones we have seen before, is very closely connected with continuous things. Suppose that 'incidents' occur at random times, but at a steady rate overall. The best example is radioactive decay: atomic nuclei decay randomly, but the average number λ which will decay in a given interval is constant. The Poisson random variable X counts the number of 'incidents' which occur in a given interval. So if, on average, there are 2.4 nuclear decays per second, then the number of decays in one second starting now is a $Poisson(2.4)$ random variable. Another example might be the number of telephone calls a minute to a busy telephone number.

The p.m.f. for a $X \sim Poisson(\lambda)$ is given by the formula

$$P(X = k) = \frac{\lambda^k}{k!} \exp^{-\lambda} \tag{26}$$

The expected value and variance of are given by

$$E(X) = Var(X) = \lambda \tag{27}$$

3.3 Continuous Random Variable

Continuous random variables, unlike discrete counterparts, take on values that vary continuously within one or more real intervals, and have a cumulative distribution function (CDF) that is absolutely continuous. As a result, the random variable has an uncountable infinite number of possible values, all of which have probability 0, though ranges of such values can have non-zero probability. **(This is counter intuitive in my opinion that in range probability exists but it is 0 for a particular point. The reason might be that, as their can be infinite real numbers in a given range, ans suppose they are ought to sum to 1 because of the fundamental property of probability, each value of a r.v. will have negligible probability hence probability at exact point is 0. This is my understanding. But again, it holds true only for the uniform distribution where we are needed to probability equally over all the points. But, in general distribution, this may not hold true, as some points can have high probability while others having 0.)** The resulting probability distribution of the random variable can be described by a probability density, where the probability is found by taking the area under the curve.

Cumulative Distribution Function (CDF) As the probability at exact point is 0 in the continuous random variable case, we can not use probability mass function as we did in discrete case; it would always be zero and give no information. Instead of that CDF is used. The c.d.f. of the random variable X is the function F X defined by

$$F_X(x) = P(X \leq x). \tag{28}$$

Note: The name of the function is F_X; the lower case x refers to the argument of the function, the number which is substituted into the function. It is common but not universal to use as the argument the lower-case version of the name of the random variable, as here. Note that $F_X(y)$ is the same function written in terms of the variable y instead of x, whereas $F_Y(x)$ is the c.d.f. of the random variable Y (which might be a completely different function.)

Now let X be a continuous random variable. Then, since the probability that X takes the precise value x is zero, there is no difference between $P(X \leq x)$ and $P(X < x)$.

The c.d.f. is an increasing function (this means that $F_X(x) \leq F_X(y)$ if $x < y$), and approaches the limits 0 as $x \to -\infty$ and 1 as $x \to \infty$.

Probability density function Another important function is the probability density function f_X. It is obtained by differentiating the c.d.f.:

$$f_X = \frac{d}{dx} F_X(x) \tag{29}$$

Now $f_X(x)$ is non-negative, since it is the derivative of an increasing function. If we know $f_X(x)$, then F_X is obtained by integrating. Because $F_X(-\infty) = 0$, we have

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \tag{30}$$

Note the use of the "dummy variable" t in this integral. Note also that

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(t)dt \tag{31}$$

We can think of the p.d.f. like this: the probability that the value of X lies in a very small interval from x to $x + h$ is approximately $f_X(x)h$. So, although the probability of getting exactly the value x is zero, the probability of being close to x is proportional to $f_X(x)$.

There is a mechanical analogy which you may find helpful. Remember that we modelled a discrete random variable X by placing at each value a of X a mass equal to P(X = a). Then the total mass is one, and the expected value of X is the centre of mass. For a continuous random variable, imagine instead a wire of variable thickness, so that the density of the wire (mass per unit length) at the point x is equal to $f_X(x)$. Then again the total mass is one; the mass to the left of x is $F_X(x)$; and again it will hold that the centre of mass is at E(X).

Most facts about continuous random variables are obtained by replacing the p.m.f. by the p.d.f. and replacing sums by integrals. Thus, the expected value of X is given by

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx \tag{32}$$

and variance is

$$Var(X) = E(X^2) - (E(X))^2 \qquad \text{where } E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)dx \tag{33}$$

Medians, quartiles, and percentiles Another measure commonly used for continuous random variables is the median; this is the value m such that "half of the distribution lies to the left of m and half to the right". More formally, m should satisfy $F_X(m) = \frac{1}{2}$, in other words, we can say that $P(X \leq m) = 0.5$. It is not the same as the mean or expected value. Except some special cases, if there is a value m such that the graph of $y = f_X(x)$ is symmetric about $x = m$, then both the expected value and the median of X are equal to m.

The lower quartile l and the upper quartile u are similarly defined by

$$F_X(l) = 1/4$$
$$F_X(u) = 3/4$$

Thus, the probability that X lies between l and u is 3/4 - 1/4 = 1/2, so the quartiles give an estimate of how spread-out the distribution is. More generally, we define the nth percentile of X to be the value of x_n such that $F_X(x_n) = n/100$.

3.4 Popular continuous probability distributions

Now, let us see some commonly used continuous variable probability distributions.

3.4.1 Uniform Distribution

Let a and b be real numbers with $a < b$. A uniform random variable on the interval $[a, b]$ is, roughly speaking, equally likely to be anywhere in the interval. In other words, its probability density function is constant on the interval $[a, b]$ (and zero outside the interval). What should the constant value c be? The integral of the p.d.f. is the area of a rectangle of height c and base $b - a$; this must be 1, so $c = 1/(b-a)$. Thus, the p.d.f. of the random variable $X \sim U(a, b)$ is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & if \ \ a \leq x \leq b \\ 0 & otherwise \end{cases}$$

By integration c.d.f. is given as

$$F_X(x) = \begin{cases} 0 & if \ \ x < a \\ \frac{x-a}{b-a} & if \ \ a \leq x \leq b \\ 1 & if \ \ \ x > b \end{cases}$$

Further calculation (or the symmetry of the p.d.f.) shows that the expected value and the median of X are both given by $(a + b)/2$ (the midpoint of the interval), while $Var(X) = (b - a)^2/12$.

3.4.2 Exponential random variable

The exponential random variable arises in the same situation as the Poisson: be careful not to confuse them! We have events which occur randomly but at a constant average rate of λ per unit time (e.g. radioactive decays, fish

biting). The Poisson random variable, which is discrete, counts how many events will occur in the next unit of time. The exponential random variable, which is continuous, measures exactly how long from now it is until the next event occurs. Note that it takes non-negative real numbers as values.

If $X \sim Exp(\lambda)$, the p.d.f. of X is

$$f_X(x) = \begin{cases} 0 & if \ x < 0 \\ \lambda \exp^{-\lambda x} & if \ x \geq 0 \end{cases}$$

CDF is given as

$$F_X(x) = \begin{cases} 0 & if \ x < 0 \\ 1 - \exp^{-\lambda x} & if \ x \geq 0 \end{cases}$$

Further calculation gives

$$E(X) = \frac{1}{\lambda}, \qquad Var(X) = \frac{1}{\lambda^2}.$$

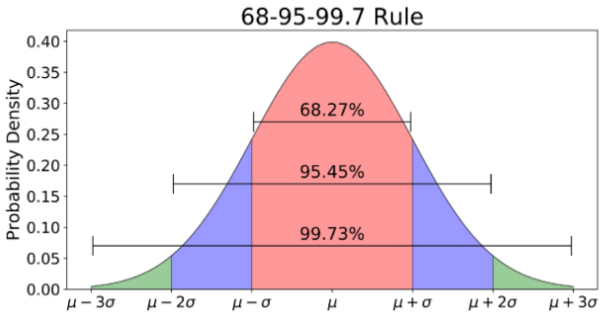

Figure 1: Normal distribution, bell shaped curve for mean value μ. It shows range of standard deviations.

3.4.3 Laplace Distribution

A closely related probability distribution that allows us to place a sharp peak of probability mass at an arbitrary point μ is the Laplace distribution

$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\Big(- \frac{|x - \mu|}{\gamma} \Big) \tag{34}$$

3.4.4 Gaussian Distribution

The Gaussian random variable is the commonest of all in applications, and the most important. There is a theorem called the central limit theorem which says that, for virtually any random variable X which is not too bizarre, if you

take the sum (or the average) of n independent random variables with the same distribution as X, the result will be approximately normal, and will become more and more like a normal variable as n grows. This partly explains why a random variable affected by many independent factors, like a man's height, has an approximately normal distribution.

More precisely, if n is large, then a $Bin(n, p)$ random variable is well approximated by a normal random variable with the same expected value np and the same variance $np(1-p)$.

The p.d.f. of the random variable $X \sim N(\mu, \sigma^2)$ is given by the formula

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{35}$$

We have $E(X) = \mu$ and $Var(X) = \sigma^2$. Figure 1 shows the graph of this function, the familiar bell-shaped curve.

When we evaluate the PDF, we need to square and invert σ. When we need to frequently evaluate the PDF with different parameter values, a more efficient way of parametrizing the distribution is to use a parameter $\beta \in (0, \infty)$ to control the precision or inverse variance of the distribution:

$$f_X(x) = \sqrt{\frac{\beta}{2\pi}} \exp^{-\frac{\beta(x-\mu)^2}{2}} \tag{36}$$

One important property of normal distribution is that, if $X \sim N(\mu, \sigma^2)$ and $Y = \frac{X-\mu}{\sigma}$ then, $Y \sim N(0, 1)$. This r.v. Y is called *standard normal variable.*

Maximum Likelihood Estimation of μ and σ Suppose, we observe n values of the data as $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$. We assume that the distribution is normal and parameters of the distribution are μ, σ. Now, with the use of estimation techniques, we intend to predict these values.

Likelihood of observing such values if we consider they are obtained independently (and of course with identical distribution), is given as:

$$p(\mathbf{x}|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{37}$$

Taking log of this likelihood function:

$$\begin{aligned} \log p(\mathbf{x}|\mu, \sigma) &= \sum_{i=1}^{n} \log \big(\frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \big) \\ &= -n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

Now, suppose we want to maximize the likelihood function in the maximum likelihood estimation, then we take derivative of the function w.r.t μ and σ and equate it to 0 to get the value of μ and σ.

Differentiating w.r.t μ

$$\begin{aligned}\frac{\partial \log p(\mathbf{x}|\mu,\sigma)}{\partial \mu} &= -\frac{\partial}{\partial \mu}(n\log(\sigma\sqrt{2\pi}) + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-\mu)(-1) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\end{aligned}$$

Equating the above equation to 0, we get

$$\mu = \frac{1}{n}\sum_{i=1}^{n}x_i \tag{38}$$

Now, we differentiate the log-likelihood function w.r.t. σ

$$\begin{aligned}\frac{\partial \log p(\mathbf{x}|\mu,\sigma)}{\partial \sigma} &= -\frac{\partial}{\partial \sigma}(n\log(\sigma\sqrt{2\pi}) + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}) \\ &= -n\frac{1}{\sigma\sqrt{2\pi}}\sqrt{2\pi} + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^3} \\ 0 &= -n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ n &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned} \tag{39}$$

We see, that the last equation produces maximum likelihood estimation for σ^2. Note that, in that equation we need the value of μ which is unknown, therefore we use the estimated value μ_{MLE}. Thus, we can do a sequential estimation, where we first estimate mean and then use that to estimate σ.

Normal distributions are a sensible choice for many applications. In the absence of prior knowledge about what form a distribution over the real numbers should take, the normal distribution is a good default choice for two major reasons.

First, many distributions we wish to model are truly close to being normal distributions. The central limit theorem shows that the sum of many independent random variables is approximately normally distributed. This means that in practice, many complicated systems can be modelled successfully as normally distributed noise, even if the system can be decomposed into parts with more structured behaviour.

Second, out of all possible probability distributions with the same variance, the normal distribution encodes the maximum amount of uncertainty over the real numbers. We can thus think of the normal distribution as being the one that inserts the least amount of prior knowledge into a model. Fully developing and justifying this idea requires more mathematical tools, and is postponed to section.

Multivariate Normal Distribution So far, we looked at the distribution where random variables were taking single value. Suppose, random variables can take vectors as input, in other words we can consider each element of vector as a value of distinct random variables which are also distributed normally. These random variables combinely produce multivariate random distribution whose density function is given as:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\} \tag{40}$$

In the above equation, $\mathbf{x}$ is an instance of vector where $x_1, x_2, \cdots, x_D$ are the certain values which are also derived from a normal distribution with mean $\mu_1, \mu_2, \cdots, \mu_D$, these mean values are compactly presented with a column vector $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_D]^T$ and their covariance matrix is given as $\mathbf{\Sigma}$.

Interpretation of $f_{\mathbf{X}}(\mathbf{x})$ is that, for a random variable $\mathbf{X}$ which takes vector values, for a certain value of vector $\mathbf{x}$, the function $f_{\mathbf{X}}(\mathbf{x})$ produces probability.

3.5 Moment Generating Function

The moment-generating function of a real-valued random variable is an alternative specification of its probability distribution. Thus, it provides the basis of an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions. There are particularly simple results for the moment-generating functions of distributions defined by the weighted sums of random variables. However, not all random variables have moment-generating functions.

As its name implies, the moment generating function can be used to compute a distribution's moments: the n^{th} moment about 0 is the n^{th} derivative of the moment-generating function, evaluated at 0.

In addition to real-valued distributions (univariate distributions), moment-generating functions can be defined for vector or matrix-valued random variables, and can even be extended to more general cases.

Let X be a random variable with cdf F_X. The moment generating function (mgf) of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = E(e^{tX}) \tag{41}$$

provided this expectation exists for t in some neighborhood of 0. That is, there is an $h > 0$ such that $\forall t \in -h < t < h$, $E\left[e^{tX}\right]$ exists. If the expectation does

not exist in a neighborhood of 0, we say that the moment generating function does not exist.

The moment-generating function is so named because it can be used to find the moments of the distribution. This can be seen by expanding exponential function as:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \frac{t^4X^4}{4!} + \cdots$$

Putting this in the expectation:

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= E(1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \frac{t^4X^4}{4!} + \cdots) \\ &= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \frac{t^4}{4!}E(X^4) + \cdots + \frac{t^n}{n!}E(X^n) + \cdots \end{aligned}$$

where $E(X^n)$ is the n^{th} moment. Differentiating $M_X(t)$ i times with respect to t and setting t=0, we obtain the i^{th} moment about the origin, $E(X^i)$. This is possible because if we look at the above expansion, terms greater than i will have t term in them after the differentiation with t, then if we put $t = 0$ all these higher order terms will become 0. Also, in the lower order terms they will no longer remain function of t hence are treated as constants and their differentiation will also become 0, as a result, we get only the term which is at i. For example, suppose we want $E(X^3)$, then as per the rule we have to differentiate thrice. When we first differentiate, 1 will be removed, and tX term will become constant X and tX^2 will be $2tX$. The second differentiation will remove X and will make $2tX \to X$. In third differentiation with t this will be also removed and only $E(X^3)$ will remain for < 3 (also note that constant of 3×2 will also be adjusted because of 3! term). For the terms > 3, there be term t in them so, once we substitute $t = 0$ these terms will become 0, giving only value for $E(X^3)$.

4 Multiple Random Variables

We have seen the joint p.m.f. of two discrete random variables X and Y, and we have learned what it means for X and Y to be independent. Now we examine this further to see measures of non-independence and conditional distributions of random variables.

4.1 Covariance and correlation

In this section we consider a pair of discrete random variables X and Y. Remember that X and Y are independent if $P(X = a_i, Y = b_j) = P(X = a_i)P(Y = b_j)$ holds for any pair (a_i, b_j) of values of X and Y. We introduce a number called the covariance of X and Y which gives a measure of how far they are from being independent.

We stated that if X and Y are independent then $Var(X+Y) = Var(X) + Var(Y)$. We found that, in any case,

$$Var(X+Y) = Var(X) + Var(Y) + 2(E(XY) - E(X)E(Y))$$

, and then proved that if X and Y are independent then $E(XY) = E(X)E(Y)$, so that the last term is zero.

Now we define the covariance of X and Y

$$Cov(X,Y) = E(XY) - E(X)E(Y) \tag{42}$$

Thus, in general for rvs X and Y,

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) \tag{43}$$

From the definition of covariance we can easily write,

$$Var(aX+bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$$

Another quantity closely related to covariance is correlation coefficient, is given as:

$$corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{44}$$

The value of correlation coefficient always lies in the range of $[-1,1]$.

4.2 Conditional random variables

Remember that the conditional probability of event B given event A is $P(B|A) = P(A \cap B)/P(A)$. Suppose that X is a discrete random variable. Then the conditional probability that X takes a certain value a_i, given A, is just

$$P(X = a_i|A) = \frac{P(\text{A holds and } X = a_i)}{P(A)}$$

Now the event A might itself be defined by a random variable. Then for such dependence we can conditional expectation as:

$$E(X|Y = b_j) = \sum_i a_i P(X = a_i|Y = b_j) \tag{45}$$

Also from this, we can write,

$$E(X) = \sum_j E(X|Y = b_j)P(Y = b_j) \tag{46}$$

4.3 Joint Distribution

Let X and Y be continuous random variables. The joint cumulative distribution function of X and Y is the function $F_{X,Y}$ of two real variables given by

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) \tag{47}$$

The joint probability density function of X and Y is

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \, \partial y} F_{X,Y}(x, y) \tag{48}$$

The probability that the pair of values of (X,Y) corresponds to a point in some region of the plane is obtained by taking the double integral of $f_{X,Y}$ over that region. For example,

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_b^a f_{X,Y}(x, y) dx \, dy$$

Marginal pdf of X and Y are given as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \qquad\qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$$

The expected value of XY is, not surprisingly,

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x, y) dx \, dy \tag{49}$$

5 Transformation of Random Variable

It may seem that there are very many known distributions, but in reality the set of distributions for which we have names is quite limited. Therefore, it is often useful to understand how transformed random variables are distributed. For example, assuming that X is a random variable distributed according to the univariate normal distribution $\mathcal{N}(0, 1)$, what is the distribution of X^2? Another example, which is quite common in machine learning, is, given that X_1 and X_2 are univariate standard normal, what is the distribution of $(X_1 + X_2)$? One option to work out the distribution of $(X_1 + X_2)$ is to calculate the mean and variance of X_1 and X_2 and then combine them. We can calculate the mean and variance of resulting random variables when we consider affine transformations of random variables. However, we may not be able to obtain the functional form of the distribution under transformations. Furthermore, we may be interested in nonlinear transformations of random variables for which closed-form expressions are not readily available.

We will look at two approaches for obtaining distributions of transformations of random variables: a direct approach using the definition of a cumulative distribution function and a change-of-variable approach that uses the chain rule

of calculus. The change-of-variable approach is widely used because it provides a "recipe" for attempting to compute the resulting distribution due to a transformation. We will explain the techniques for univariate random variables, and will only briefly provide the results for the general case of multivariate random variables. Transformations of discrete random variables can be understood directly. Suppose that there is a discrete random variable X with pmf $P(X = x)$, and an invertible function $U(x)$. Consider the transformed random variable $Y \coloneqq U(X)$, with pmf $P(Y = y)$. Then

$$P(Y = y) = P(U(X) = y) = P(X = U^{-1}(y)) \tag{50}$$

transformation of interest inverse where we can observe that $x = U^{-1}(y)$. Therefore, for discrete random variables, transformations directly change the individual events (with the probabilities appropriately transformed).

Let X be a continuous random variable. Let g be a real function which is either strictly increasing or strictly decreasing on the support of X, and which is differentiable there. Let $Y = g(X)$. Then (a) the support of Y is the image of the support of X under g; (b) the p.d.f. of Y is given by $f_Y(y) = f_X(h(y))|h'(y)|$, where h is the inverse function of g.

5.1 Distribution Function Technique

The distribution function technique goes back to first principles, and uses the definition of a cdf $F_X(x) = P(X \leq x)$ and the fact that its differential is the pdf $f(x)$. For a random variable X and a function U, we find the pdf of the random variable $Y \coloneqq U(X)$ by

1. Finding the cdf:
$$F_Y(y) = P(Y \leq y) \tag{51}$$
To obtain this cdf, we start with the cdf of X, then use the $Y = U(X)$ relation which is given finally make the appropriate substations to get $F_Y(y)$.

2. Differentiating the cdf $F_Y(y)$ to get the pdf $f(y)$.

Consider this example, let X be a continuous random variable with probability density function $0 \leq x \leq 1$, and pdf $f(x) = 3x^2$. The new random variable Y is related to X as $Y = X^2$. Then find out Y's pdf.

First we obtain cdf of Y from X as,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(X^2 \leq y) \\ &= P(X \leq y^{\frac{1}{2}}) \\ &= F_X(y^{\frac{1}{2}}) \\ &= \int_0^{y^{\frac{1}{2}}} 3t^2 dt \\ &= [t^3]_0^{y^{\frac{1}{2}}} \\ F_Y(y) &= y^{\frac{3}{2}}, \qquad 0 \leq y \leq 1 \end{aligned}$$

Now, pdf is obtained by differentiating it,

$$f(y) = \frac{d}{dy} F_Y(y) = \frac{3}{2} y^{\frac{1}{2}} \qquad 0 \leq y \leq 1$$

5.2 Change of Variables

The previous technique of distribution function is derived from first principles, based on the definitions of cdfs and using properties of inverses, differentiation, and integration. This argument from first principles relies on two facts:

1. We can transform the cdf of Y into an expression that is a cdf of X.
2. We can differentiate the cdf to obtain the pdf.

Let us break down the reasoning step by step, with the goal of understanding the more general change-of-variables approach.

Consider a univariate random variable X, and an invertible function U, which gives us another random variable $Y = U(X)$. We assume that random variable X has states $x \in [a, b]$. By the definition of the cdf, we

$$F_Y(y) = P(Y \leq y) = P(U(X) \leq y)$$

In the case $U(X)$ is strictly increasing, $U^{-1}(X)$ is also strictly increasing, therefore we can write

$$P(U(X) \leq y) = P(U^{-1}(U(X)) \leq U^{-1}(y)) = P(X \leq U^{-1}(y))$$

With the definitions of pdf and cdf, we can write the followings, which we have already used in the previous sections,

$$F_Y(y) = P(X \leq U^{-1}(y)) = \int_a^{U^{-1}(y)} f(x) dx$$

To obtain pdf, we differentiate the cdf

$$f(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} \int_a^{U^{-1}(y)} f(x)dx \tag{52}$$

Note that the integral on the right-hand side is with respect to x, but we need an integral with respect to y because we are differentiating with respect to y. In particular, we make the integration rule to get the substitution

$$\int f(x)dx = \int f(U^{-1}(y))U'^{-1}(y)dy \qquad \text{where } x = U^{-1}(y)$$

with this substitution, we can write the following

$$\begin{aligned} f(y) &= \frac{d}{dy} \int_a^{U^{-1}(y)} f_x(U^{-1}(y))U'^{-1}(y)dy \\ &= f_x(U^{-1}(y))U'^{-1}(y) \qquad \text{derivative and integration can be cancelled} \\ &= f_x(U^{-1}(y)) \frac{d}{dy} U^{-1}(y) \qquad \text{simple arrangements} \\ f(y) &= f_x(U^{-1}(y)) |\frac{d}{dy} U^{-1}(y)| \end{aligned}$$

In the last step, we introduce absolute as we may get negative derivative if the function U is decreasing **but then how can we use the decreasing function in the previous steps?**. This is called the change-of-variable technique. The term $|\frac{d}{dy} U^{-1}(y)|$ measures how much a unit volume changes when applying U.

By change of variable technique the pdf of new variable $Y = U(X)$ if Y is differentiable and either increasing or decreasing is given as

$$f(y) = f_x(U^{-1}(y)) |\frac{d}{dy} U^{-1}(y)| \tag{53}$$

Note that, in comparison to the discrete case in Eq. 50, we have an additional factor $|\frac{d}{dy} U^{-1}(y)|$. The continuous case requires more care because $P(Y = y) = 0 \ \forall y$. The probability density function $f(y)$ does not have a description as a probability of an event involving y.

5.3 Discussion on Summation vs Multiplication

There is possibility of summing two random variables, this operation produces new random variable. We want to discuss, what exactly happens in that case with an example.

Suppose there is a random variable $X \sim P_X$ with some probability distribution and another random variable $Y \sim P_Y$ with different variable. These both

probability distributions are over the experiment of throwing a dice and take values corresponding to the output of the die i.e. $\{1, 2, \cdots, 6\}$. The die is not fair and the probabilities are assigned unevenly. The probability mass function of these two distributions is shown in Figure 2.

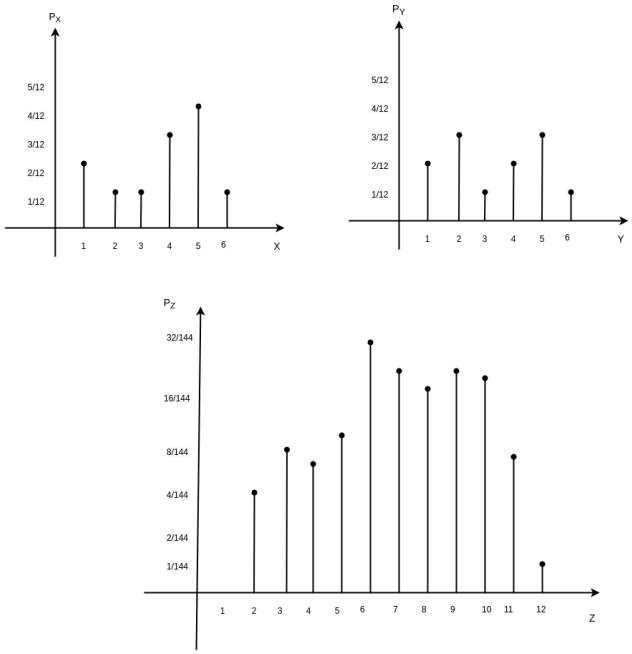

Figure 2: Probability mass functions of X and Y random variables. They are are added to produce pmf for $Z = X + Y$.

The random variables X and Y are added to produce new r.v. which has range from $2, 3, \cdots, 12$. Then for $Z \sim P_Z$ is obtained by combining those events which produce these sums. For example, the sum of 6 is produced with 5 different event combinations from X and Y, $\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$. Then to get the probability of $Z = 6$, probability of all these events is summed. The pmf of Z is shown in figure to indicate this.

This differs from the multiplication of random variables, where the range of the r.v. will be from 1 to 144, and it will give different set of events than the summation. But, the calculation of the probabilities will be similar. For example, 2 from multiplication will be obtained with $\{(1, 2), (2, 1)\}$, so probabilities of these will be summed to get the probability for the multiplication 2.

One thing is common in both the cases is joint probability of the two random variables while getting the probability of the new variable. Consider the situation where we want to get the probability of $(1, 2)$, it is nothing but the joint probability $P_X Y(X = 1, Y = 2)$. The space for the join probability is the cross product between *ranges* of X and Y, i.e. $\{1, 2, \cdots, 6\} \times \{1, 2, \cdots, 6\}$ which

will be $\{(1,1),(1,2)\cdots,(6,6)\}$ containing 36 elements. This is different than the summation or the multiplication random variable. But implicitly we have to know this joint distributions to obtain the distribution of both summation and multiplication of random variables.