1. Probability Theory

Probability theory is a branch of pure mathematics, and forms the theoretical basis of statistics. In itself, probability theory has some basic objects and their relations (like real numbers, addition, etc. for analysis) and it makes no pretense of saying anything about the real world. Axioms are given and theorems are then deduced about these objects, just as in any other part of mathematics.

But, a very important aspect of probability is that it is *applicable*. In other words, there are many real world situations in which it is reasonable to take a model in probability and it turns out to reasonably replicate features of the real world situation.

2. Discrete probability spaces

Definition 1. Let Ω be a finite or countable set. Let $p : \Omega \to [0,1]$ be a function such that

$$\sum_{\omega \in \Omega} p_\omega = 1.$$

Then, (Ω, p) is called a *discrete probability space*. Ω is called the *sample space* and p_ω are called *elementary probabilities*.

- Any subset $A \subseteq \Omega$ is called an *event*. For an event A, we define its *probability* as $\mathbf{P}(A) = \sum_{\omega \in A} p_\omega$.
- Any function $X : \Omega \to \mathbb{R}$ is called a *random variable*.

All of probability in one line: Take an (interesting) probability space (Ω, p) and an (interesting) event $A \subseteq \Omega$. Find $\mathbf{P}(A)$.

This is the mathematical side of the picture. It is easy to make up any number of probability spaces - simply take a finite set and assign non-negative numbers to each element of the set so that the total is 1.

Example 2. $\Omega = \{0,1\}$ and $p_0 = p_1 = \frac{1}{2}$. There are only four events here, $\emptyset, \{0\}, \{1\}$ and $\{0,1\}$. Their probabilities are, 0, 1/2, 1/2 and 1, respectively.

Example 3. $\Omega = \{0,1\}$. Fix a number $0 \le p \le 1$ and let $p_1 = p$ and $p_0 = 1 - p$. The sample space is the same as before, but the probability space is different for each value of p. Again, there are only four events, and their probabilities are $\mathbf{P}\{\emptyset\} = 0$, $\mathbf{P}\{0\} = 1 - p$, $\mathbf{P}\{1\} = p$ and $\mathbf{P}\{0,1\} = 1$.

Example 4. Fix a positive integer n. Let

$$\Omega = \{0,1\}^n = \{\underline{\omega} : \underline{\omega} = (\omega_1, \ldots, \omega_n) \text{ with } \omega_i = 0 \text{ or } 1 \text{ for each } 1 \le i \le n\}.$$

Let $p_{\underline{\omega}} = 2^{-n}$ for each $\underline{\omega} \in \Omega$. Since Ω has 2^n elements, it follows that this is a valid assignment of elementary probabilities.

There are $2^{\#\Omega} = 2^{2^n}$ events [1]. One example is $A_k = \{\underline{\omega} : \underline{\omega} \in \Omega \text{ and } \omega_1 + \cdots + \omega_n = k\}$, where k is some fixed integer. In words, A_k consists of those n-tuples of zeros and ones that have a total of k many ones. Since there are $\binom{n}{k}$ ways to choose where to place these ones, we see that $\#A_k = \binom{n}{k}$. Consequently,

$$\mathbf{P}(A_k) = \sum_{\underline{\omega} \in A_k} p_{\underline{\omega}} = \frac{\#A_k}{2^n} = \begin{cases} \binom{n}{k} 2^{-n} & \text{if } 0 \le k \le n, \\ 0 & \text{otherwise.} \end{cases}$$

It will be convenient to adopt the notation that $\binom{a}{b} = 0$ if a, b are positive integers and if $b > a$ or if $b < 0$. Then, we can simply write $\mathbf{P}(A_k) = \binom{n}{k} 2^{-n}$ without having to split the values of k into cases.

[1] $\#\Omega$ denotes the cardinality of the set Ω.

Example 5. Fix two positive integers r and m. Let

$$\Omega = \{\underline{\omega} : \underline{\omega} = (\omega_1, \ldots, \omega_r) \text{ with } 1 \le \omega_i \le m \text{ for each } 1 \le i \le r\}.$$

The cardinality of Ω is m^r (since each co-ordinate ω_i can take one of m values). Hence, if we set $p_{\underline{\omega}} = m^{-r}$ for each $\underline{\omega} \in \Omega$, we get a valid probability space.

Of course, there are 2^{m^r} many events, which is quite large even for small numbers like $m = 3$ and $r = 4$. Some interesting events are $A = \{\underline{\omega} : \omega_r = 1\}$, $B = \{\underline{\omega} : \omega_i \neq 1 \text{ for all } i\}$, $C = \{\underline{\omega} : \omega_i \neq \omega_j \text{ if } i \neq j\}$. The reason why these are interesting will be explained later. Because of equal elementary probabilities, the probability of an event S is just $\#S/m^r$.

- Counting A: We have m choices for each of $\omega_1, \ldots, \omega_{r-1}$. There is only one choice for ω_r. Hence $\#A = m^{r-1}$. Thus, $\mathbf{P}(A) = \frac{m^{r-1}}{m^r} = \frac{1}{m}$.
- Counting B: We have $(m-1)$ choices for each ω_i (since ω_i cannot be 1). Hence $\#B = (m-1)^r$ and thus $\mathbf{P}(B) = \frac{(m-1)^r}{m^r} = (1 - \frac{1}{m})^r$.
- Counting C: We must choose a distinct value for each $\omega_1, \ldots, \omega_r$. This is impossible if $m < r$. If $m \ge r$, then ω_1 can be chosen as any of m values. After ω_1 is chosen, there are $(m-1)$ possible values for ω_2, and then $(m-2)$ values for ω_3 etc., all the way till ω_r which has $(m-r+1)$ choices. Thus, $\#C = m(m-1)\cdots(m-r+1)$. Note that we get the same answer if we choose ω_i in a different order (it would be strange if we did not!). Thus, $\mathbf{P}(C) = \frac{m(m-1)\cdots(m-r+1)}{m^r}$.

2.1. **Probability in the real world.** In real life, there are often situations where there are several possible outcomes but which one will occur is unpredictable in some way. For example, when we toss a coin, we may get heads or tails. In such cases we use words such as *probability or chance*, *event or happening*, *randomness*, etc. What is the relationship between the intuitive and mathematical meanings of words such as probability or chance?

In a given physical situation, we choose one out of all possible probability spaces that we think captures best the chance happenings in the situation. The chosen probability space is then called a *model* or a *probability model* for the given situation. Once the model has been chosen, calculation of probabilities of events therein is a mathematical problem. Whether the model really captures the given situation, or whether the model is inadequate and over-simplified is a non-mathematical question. Nevertheless that is an important question, and can be answered by observing the real life situation and comparing the outcomes with predictions made using the model[2].

Now we describe several "random experiments" (a non-mathematical term to indicate a "real-life" phenomenon that is supposed to involve chance happenings) in which the previously given

[2] Roughly speaking we may divide the course into two parts according to these two issues. In the probability part of the course, we shall take many such models for granted and learn how to calculate (or, approximately calculate) probabilities. In the statistics part of the course, we shall see some methods by which we can arrive at such models, or test the validity of a proposed model.

examples of probability spaces arise. Describing the probability space is the first step in any probability problem.

Example 6. Physical situation: Toss a coin. Randomness enters because we believe that the coin may turn up head or tail and that it is inherently unpredictable.

The corresponding probability model: Since there are two outcomes, the sample space $\Omega = \{0, 1\}$ (where we use 1 for heads and 2 for tails) is a clear choice. What about elementary probabilities? Under the equal chance hypothesis, we may take $p_0 = p_1 = \frac{1}{2}$. Then, we have a probability model for the coin toss.

If the coin was not fair, we would change the model by keeping $\Omega = \{0, 1\}$ as before but letting $p_1 = p$ and $p_0 = 1 - p$ where the parameter $p \in [0, 1]$ is fixed.

Which model is correct? If the coin looks symmetrical, then the two sides are equally likely to turn up, so the first model where $p_1 = p_0 = \frac{1}{2}$ is reasonable. However, if the coin looks irregular, then theoretical considerations are usually inadequate to arrive at the value of p. Experimenting with the coin (by tossing it a large number of times) is the only way.

There is always an approximation in going from the real-world to a mathematical model. For example, the model above ignores the possibility that the coin can land on its side. If the coin is very thick, then it might be closer to a cylinder which can land in three ways and then we would have to modify the model.

Thus, we see that Example 3 is a good model for a physical coin toss. What physical situations are captured by the probability spaces in Example 4 and Example 5?

Example 4: This probability space can be a model for tossing n fair coins. It is clear in what sense, so we omit details for you to fill in.

The same probability space can also be a model for the tossing of the same coin n times in succession. In this, we are implicitly assuming that the coin forgets the outcomes on the previous tosses. While that may seem obvious, it would be violated if our "coin" was a hollow lens filled with a semi-solid material like glue (then, depending on which way the coin fell on the first toss, the glue would settle more on the lower side and consequently the coin would be more likely to fall the same way again). This is a coin with memory!

Example 5: There are several situations that can be captured by this probability space. We list some.

- There are r labelled balls and m labelled bins. One by one, we put the balls into bins "at random". Then, by letting ω_i be the bin-number into which the i^{th} ball goes, we can capture the full configuration by the vector $\underline{\omega} = (\omega_1, \ldots, \omega_n)$. If each ball is placed completely at random then the probabilities are m^{-r} for each configuration $\underline{\omega}$.

In that example, A is the event that the last ball ends up in the first bin, B is the event that the first bin is empty and C is the event that no bin contains more than one ball.

- If $m = 6$, then this may also be the model for throwing a fair die r times. Then ω_i is the outcome on the i^{th} throw. Of course, it also models throwing r different (and distinguishable) fair dice.
- If $m = 2$ and $r = n$, this is same as Example 4, and thus models the tossing of n fair coins (or a fair coin n times).
- Let $m = 365$. Omitting the possibility of leap years, this is a model for choosing r people at random and noting their birthdays (which can be in any of 365 "bins"). If we assume that all days are equally likely as a birthday (is this really true?), then the same probability space is a model for this physical situation. In this example, C is the event that no two people have the same birthday.

Example 7. Toss n coins. We saw this before, but assumed that the coins are fair. Now we do not. The sample space is

$$\Omega = \{0,1\}^n = \{\underline{\omega} = (\omega_1, \ldots, \omega_n) : \omega_i = 0 \text{ or } 1 \text{ for each } i \le n\}.$$

Further, we assign $p_{\underline{\omega}} = \alpha^{(1)}_{\omega_1} \ldots \alpha^{(n)}_{\omega_n}$. Here, $\alpha^{(j)}_0$ and $\alpha^{(j)}_1$ are supposed to indicate the probabilities that the j^{th} coin falls tails up or heads up, respectively. Why did we take the product of $\alpha^{(j)}_{\cdot}$s and not some other combination? This is a non-mathematical question about what model is suited for the given real-life example. For now, the only justification is that empirically the above model seems to capture the real life situation accurately.

In particular, if the n coins are identical, we may write $p = \alpha^{(j)}_1$ (for any j) and the elementary probabilities become $p_{\underline{\omega}} = p^{\sum_i \omega_i} q^{n - \sum_i \omega_i}$, where $q = 1 - p$.

Fix $0 \le k \le n$ and let $B_k = \{\underline{\omega} : \sum_{i=1}^n \omega_i = k\}$ be the event that we see exactly k heads out of n tosses. Then, $\mathbf{P}(B_k) = \binom{n}{k} p^k q^{n-k}$. If A_k is the event that there are at least k heads, then $\mathbf{P}(A_k) = \sum\limits_{\ell=k}^{n} \binom{n}{\ell} p^\ell q^{n-\ell}$.

Example 8. Toss a coin n times. Again

$$\Omega = \{0,1\}^n = \{\underline{\omega} = (\omega_1, \ldots, \omega_n) : \omega_i = 0 \text{ or } 1 \text{ for each } i \le n\},$$

$$p_{\underline{\omega}} = p^{\sum_i \omega_i} q^{n - \sum_i \omega_i}.$$

This is the same probability space that we got for the tossing of n identical looking coins. Implicit is the assumption that once a coin is tossed, for the next toss it is as good as a different coin but with the same p. It is possible to imagine a world where coins retain the memory of what happened before (or as explained before, we can make a "coin" that remembers previous tosses!), in which case this would not be a good model for the given situation. We don't believe that this is the case for coins in our world, and this can be verified empirically.

Example 9. Shuffle a deck of 52 cards. $\Omega = S_{52}$, the set of all permutations[3] of $[52]$ and $p_\pi = \frac{1}{52!}$ for each $\pi \in S_{52}$. More generally, we can make a model for a deck of n cards, in which case the sample space is S_n and elementary probabilities are $1/n!$ for each n.

[3] We use the notation $[n]$ to denote the set $\{1, 2, \ldots, n\}$. A permutation of $[n]$ is a vector $(i_1, i_2, \ldots, i_n)$, where $i_1, \ldots, i_n$ are distinct elements of $[n]$, in other words, they are $1, 2, \ldots, n$ but in some order. Mathematically, we may define a permutation as a bijection $\pi : [n] \to [n]$. Indeed, for a bijection π, the numbers $\pi(1), \ldots, \pi(n)$ are just $1, 2, \ldots, n$ in some order.

As an illustration, when $n = 3$, the sample space is

$$S_3 = \{(1,2,3),(1,3,2),(2,3,1),(2,1,3),(3,1,2),(3,2,1)\},$$

where $(2,3,1)$ denotes the deck where the top card is 2, the next one is 3 and the bottom card is 1. The elementary probabilities are all $1/6$ in this case.

Example 10. Toss a coin till a head turns up. $\Omega = \{1, 01, 001, 0001, \dots\} \cup \{\bar{0}\}$. Let us write $0^k1 = 0\dots01$ as a short form for k zeros (tails) followed by 1 and $\bar{0}$ stands for the sequence of all tails. Let $p \in [0,1]$. Then, we set $p_{0^k1} = q^k p$ for each $k \in \mathbb{N}$. We also set $p_{\bar{0}} = 0$ if $p > 0$ and $p_{\bar{0}} = 1$ if $p = 0$. This is forced on us by the requirement that elementary probabilities add to 1.

Let $A = \{0^k1 : k \geq n\}$ be the event that at least n tails fall before a head turns up. Then $\mathbf{P}(A) = q^n p + q^{n+1}p + \dots = q^n$.

Example 11. Place r distinguishable balls in m distinguishable urns at random. We saw this before (the words "labelled" and "distinguishable" mean the same thing here). The sample space is $\Omega = [m]^r = \{\underline{\omega} = (\omega_1, \dots, \omega_r) : 1 \leq \omega_i \leq m\}$ and $p_{\underline{\omega}} = m^{-r}$ for every $\underline{\omega} \in \Omega$. Here, ω_i indicates the urn number into which the i^{th} ball goes.

Example 12. Birthday "paradox" There are n people at a party. What is the chance that two of them have the same birthday?

This can be thought of as a balls in bin problem, where the bins are labelled $1, 2, \dots, 365$ (days of the year), and the balls are labelled $1, 2, \dots, n$ (people). The sample space can be copied from the previous example with $r = n$ and $m = 365$. In doing this, we have omitted the possibility of February 29th, which simplifies our life a little bit. What are the elementary probabilities? Let us assume that all sample points have equal probability[4].

The event of interest is that there is at least one bin that has at least two balls. In other words,

$$A = \{\underline{\omega} = (\omega_1, \dots, \omega_n) : \omega_i = \omega_j \text{ for some } i < j\}.$$

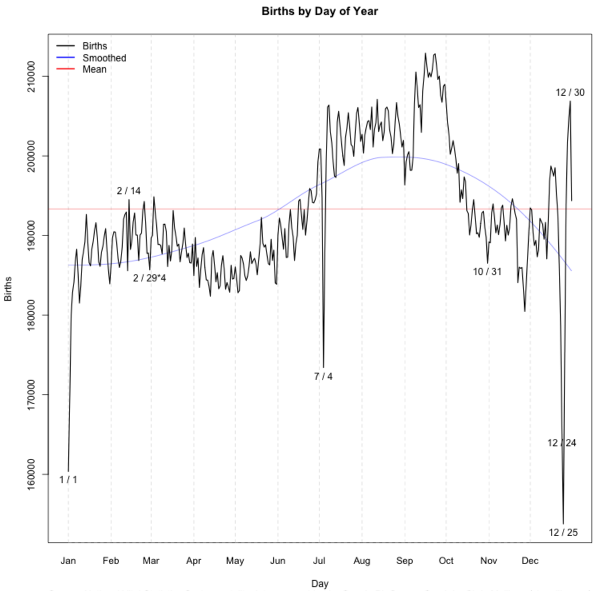
In this case, it is easier to calculate the probability of

$$A^c = \{\underline{\omega} = (\omega_1, \dots, \omega_n) : \omega_i \neq \omega_j \text{ for all } i < j\}.$$

Indeed, the cardinality of A^c is $m(m-1)\cdots(n-m+1)$ (each person has to be assigned a different birthday), whence

$$\mathbf{P}(A) = 1 - \frac{m(m-1)\cdots(m-n+1)}{m^r}.$$

[4] This assumption is not entirely realistic, for example, Figure 3 shows actual data from a specific geographic location over a specific period of time. Seasonal variation is apparent! In addition there are complications such as the possibility of a pair of twins attending the party.



Source: National Vital Statistics System natality data, as provided by Google BigQuery. Graph by Chris Mulligan (chmullig.com)

FIGURE 1. Frequencies of birthdays in the United States of America from 1969 to 1988. Data taken from Andrew Gelman.

If $n > m$, the probability is obviously ~~zero~~. The reason this is called a “paradox” is that even for n much smaller than m, the probability becomes significantly large. For example, for $m = 365$, here are the probabilities for a few values of n:

n	5	15	25	35	45
$\mathbf{P}(A)$	0.027	0.253	0.569	0.814	0.941

Example 13. Place r indistinguishable balls in m distinguishable urns at random. Since the balls are indistinguishable, we can only count the number of balls in each urn. The sample space is

$$\Omega = \{(\ell_1, \dots, \ell_m) : \ell_i \geq 0,\ \ell_1 + \cdots + \ell_m = r\}.$$

We give two proposals for the elementary probabilities.

(1) Let $p^{\text{MB}}_{(\ell_1,\dots,\ell_m)} = \frac{m!}{\ell_1!\ell_2!\cdots\ell_m!}\frac{1}{m^r}$. These are the probabilities that result if we place r labelled balls in m labelled urns, and then erase the labels on the balls.

(2) Let $p^{\text{BE}}_{(\ell_1,\dots,\ell_m)} = \frac{1}{\binom{m+r-1}{r-1}}$ for each $(\ell_1, \dots, \ell_m) \in \Omega$. Elementary probabilities are chosen so that all distinguishable configurations are equally likely.

That these are legitimate probability spaces depend on two combinatorial facts.

Notation: Let $A \subseteq \Omega$ be an event. Then, we define a function $\mathbf{1}_A : \Omega \to \mathbb{R}$, called the *indicator function of* A, as follows.

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Exercise 14. (1) Let $(\ell_1, \ldots, \ell_m) \in \Omega$. Show that $\#\{\underline{\omega} \in [m]^r : \sum_{j=1}^r \mathbf{1}_{\omega_j = i} = \ell_i \text{ for each } i \in [m]\} = \frac{n!}{\ell_1!\ell_2!\cdots\ell_m!}$. Hence or directly, show that $\sum\limits_{\omega\in\Omega} p_\omega^{MB} = 1$.

(2) Show that $\#\Omega = \binom{m+r-1}{r-1}$. Hence, $\sum\limits_{\omega\in\Omega} p_\omega^{BE} = 1$.

The two models are clearly different. Which one captures reality? We can arbitrarily label the balls for our convenience, and then erase the labels in the end. This clearly yields elementary probabilities p^{MB}. To put it another way, pick the balls one by one and assign them randomly to one of the urns. This suggests that p^{MB} is the "right one".

This leaves open the question of whether there is a natural mechanism of assigning balls to urns so that the probabilities p^{BE} shows up. No such mechanism has been found. But this probability space does occur in the physical world. If r photons ("indistinguishable balls") are to occupy m energy levels ("urns"), then empirically it has been verified that the correct probability space is the second one![5]

Example 15. Sampling with replacement from a population. Define $\Omega = \{\underline{\omega} \in [N]^k : \omega_i \in [N] \text{ for } 1 \le i \le k\}$ with $p_{\underline{\omega}} = 1/N^k$ for each $\underline{\omega} \in \Omega$. Here, $[N]$ is the population (so the size of the population is N) and the size of the sample is k. Often the language used is of a box with N coupons from which k are drawn with replacement.

Example 16. Sampling without replacement from a population. Now we take

$$\Omega = \left\{\underline{\omega} \in [N]^k : \omega_i \text{ are distinct elements of } [N]\right\},$$

$$p_{\underline{\omega}} = \frac{1}{N(N-1)\cdots(N-k+1)} \text{ for each } \underline{\omega} \in \Omega.$$

Fix $m < N$ and define the random variable $X(\underline{\omega}) = \sum_{i=1}^k \mathbf{1}_{\omega_i \le m}$. If the population $[N]$ contains a subset, say $[m]$, (could be the subset of people having a certain disease), then $X(\underline{\omega})$ counts the

[5] The probabilities p^{MB} and p^{BE} are called Maxwell-Boltzmann statistics and Bose-Einstein statistics. There is a third kind, called Fermi-Dirac statistics which is obeyed by electrons. For general $m \ge r$, the sample space is $\Omega_{\mathrm{FD}} = \{(\ell_1, \ldots, \ell_m) : \ell_i = 0 \text{ or } 1 \text{ and } \ell_1 + \cdots + \ell_m = r\}$ with equal probabilities for each element. In words, all distinguishable configurations are equally likely, with the added constraint that at most one electron can occupy each energy level.

number of people in the sample who have the disease. Using X one can define events such as $A = \{\underline{\omega} : X(\underline{\omega}) = \ell\}$ for some $\ell \leq m$. If $\underline{\omega} \in A$, then ℓ of the ω_i must be in $[m]$ and the rest in $[N] \setminus [m]$. Hence

$$\#A = \binom{k}{\ell} m(m-1)\cdots(m-\ell+1)(N-m)(N-m-1)\cdots(N-m-(k-\ell)+1).$$

As the probabilities are equal for all sample points, we get

$$\begin{aligned}\mathbf{P}(A) &= \frac{\binom{k}{\ell} m(m-1)\cdots(m-\ell+1)(N-m)(N-m-1)\cdots(N-m-(k-\ell)+1)}{N(N-1)\cdots(N-k+1)} \\ &= \frac{1}{\binom{N}{k}}\binom{m}{\ell}\binom{N-m}{k-\ell}.\end{aligned}$$

This expression arises whenever the population is subdivided into two parts and we count the number of samples that fall in one of the sub-populations.

We now give two non-examples.

Example 17. A non-example - Pick a natural number uniformly at random. The sample space is clearly $\Omega = \mathbb{N} = \{1, 2, 3, \ldots\}$. The phrase "uniformly at random" suggests that the elementary probabilities should be the same for all elements. That is $p_i = p$ for all $i \in \mathbb{N}$ for some p. If $p = 0$, then $\sum_{i\in\mathbb{N}} p_i = 0$ whereas if $p > 0$, then $\sum_{i\in\mathbb{N}} p_i = \infty$. This means that there is no way to assign elementary probabilities so that each number has the same chance to be picked.

This appears obvious, but many folklore puzzles and paradoxes in probability are based on the faulty assumption that it is possible to pick a natural number at random. For example, when asked a question like "What is the probability that a random integer is odd?", many people answer $1/2$. We want to emphasize that the probability space has to be defined first, and only then can probabilities of events be calculated. Thus, the question does not make sense to us and we do not have to answer it! [6]

Example 18. Another non-example - Throwing darts. A dart is thrown at a circular dart board. We assume that the dart does hit the board but were it hits is "random" in the same sense in which

[6] For those interested, there is one way to make sense of such questions. It is to consider a sequence of probability spaces $\Omega^{(n)} = [n]$ with elementary probabilities $p_i^{(n)} = 1/n$ for each $i \in \Omega_n$. Then, for a subset $A \subseteq \mathbb{Z}$, we consider $\mathbf{P}_n(A \cap \Omega_n) = \#(A \cap [n])/n$. If these probabilities converge to a limit x as $n \to \infty$, then we could say that A has asymptotic probability x. In this sense, the set of odd numbers does have asymptotic probability $1/2$, the set of numbers divisible by 7 has asymptotic probability $1/7$ and the set of prime numbers has asymptotic probability 0. However, this notion of asymptotic probability has many shortcomings. Many subsets of natural numbers will not have an asymptotic probability, and even sets which do have asymptotic probability fail to satisfy basic rules of probability that we shall see later. Hence, we shall keep such examples out of our system.

we say the a coin toss is random. Intuitively this appears to make sense. However our framework is not general enough to incorporate this example. Let us see why.

The dart board can be considered to be the disk $\Omega = \{(x,y) : x^2 + y^2 \leq r^2\}$ of given radius r. This is an uncountable set. We cannot assign elementary probabilities $p_{(x,y)}$ for each $(x,y) \in \Omega$ in any reasonable way. In fact the only reasonable assignment would be to set $p_{(x,y)} = 0$ for each (x,y) but then what is $\mathbf{P}(A)$ for a subset A? [7] Uncountable sums are not well defined!

We need a branch of mathematics called *measure theory* to make proper sense of uncountable probability spaces. This will not be done in this course although we shall later say a bit about the difficulties involved. The same difficulty shows up in the following "random experiments" also.

(1) **Draw a number at random from the interval** $[0,1]$. $\Omega = [0,1]$ which is uncountable.

(2) **Toss a fair coin infinitely many times**. $\Omega = \{0,1\}^{\mathbb{N}} := \{\underline{\omega} = (\omega_1, \omega_2, \ldots) : \omega_i = 0 \text{ or } 1\}$. This is again an uncountable set.

Remark 19. In one sense, the first non-example is almost irredeemable but the second non-example can be dealt with, except for technicalities beyond this course. We shall later give a set of working rules to work with such "continuous probabilities". Fully satisfactory development will have to wait for a course in measure theory.

[7] Some probability problems are geometric in nature. What this means is that our desired outcome space is in fact some area called the feasible region. To find the probability of this outcome, we find the ratio feasible region/sample space. A classic example of this type of problem is a dart board.
Let us say that $r = 10$ here. Furthermore, the bulls-eye is exactly in the center and has a 1 in diameter. If I throw a dart randomly and it hits the board, what is the probability that it hits the bulls-eye?

1. Basic rules of probability

So far we have defined the notion of probability space and probability of an event. But most often, we do not calculate probabilities from the definition. This is like in integration, where one defined the integral of a function as a limit of Riemann sums, but that definition is used only to find integrals of x^n, $\sin(x)$ and a few such functions. Instead, integrals of complicated expressions such as $x\sin(x) + 2\cos^2(x)\tan(x)$ are calculated by various rules, such as substitution rule, integration by parts, etc. In probability we need some similar rules relating probabilities of various combinations of events to the individual probabilities.

Proposition 1. *Let (Ω, p) be a discrete probability space.*

(1) For any event A, we have $0 \le \mathbf{P}(A) \le 1$. Also, $\mathbf{P}(\emptyset) = 0$ and $\mathbf{P}(\Omega) = 1$.

(2) Finite additivity of probability: *If $A_1, \dots, A_n$ are pairwise disjoint events (i.e., $A_i \cap A_j = \emptyset$ if $i \neq j$), then $\mathbf{P}(A_1 \cup \dots \cup A_n) = \mathbf{P}(A_1) + \dots + \mathbf{P}(A_n)$. In particular, $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$ for any event A.*

(3) Countable additivity of probability: *If $A_1, A_2, \dots$ is a countable collection of pairwise disjoint events, then $\mathbf{P}(\cup A_i) = \sum_i \mathbf{P}(A_i)$.*

All of these may seem obvious, and indeed they would be totally obvious if we stuck to finite sample spaces. But the sample space could be countable, and then probability of events may involve infinite sums which need special care in manipulation. Therefore we must give a proof. In writing a proof, and in many future contexts, it is useful to introduce the following notation.

Notation: Let $A \subseteq \Omega$ be an event. Then, we define a function $\mathbf{1}_A : \Omega \to \mathbb{R}$, called the *indicator function of A*, as follows.

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Since a function from Ω to $\mathbb{R}$ is called a random variable, the indicator of any event is a random variable. All information about the event A is in its indicator function (meaning, if we know the value of $\mathbf{1}_A(\omega)$, we know whether or not ω belongs to A). For example, we can write $\mathbf{P}(A) = \sum_{\omega\in\Omega} \mathbf{1}_A(\omega) p_\omega$.

Now, we prove the proposition.

Proof. (1) By definition of probability space $\mathbf{P}(\Omega) = 1$ and $\mathbf{P}(\emptyset) = 0$. If A is any event, then $\mathbf{1}_\emptyset(\omega)p_\omega \le \mathbf{1}_A(\omega)p_\omega \le \mathbf{1}_\Omega(\omega)p_\omega$. By monotonicity of sums, we get

$$\sum_{\omega\in\Omega} \mathbf{1}_\emptyset(\omega)p_\omega \le \sum_{\omega\in\Omega} \mathbf{1}_A(\omega)p_\omega \le \sum_{\omega\in\Omega} \mathbf{1}_\Omega(\omega)p_\omega.$$

As observed earlier, these sums are just $\mathbf{P}(\emptyset)$, $\mathbf{P}(A)$ and $\mathbf{P}(\Omega)$, respectively. Thus, $0 \leq \mathbf{P}(A) \leq 1$.

(2) It suffices to prove it for two sets (Why?). Let A, B be two events such that $A \cap B = \emptyset$. Let $f(\omega) = p_\omega \mathbf{1}_A(\omega)$ and $g(\omega) = p_\omega \mathbf{1}_B(\omega)$ and $h(\omega) = p_\omega \mathbf{1}_{A\cup B}(\omega)$. Then, the disjointness of A and B implies that $f(\omega) + g(\omega) = h(\omega)$ for all $\omega \in \Omega$. Thus, by linearity of sums, we get

$$\sum_{\omega\in\Omega} f(\omega) + \sum_{\omega\in\Omega} g(\omega) = \sum_{\omega\in\Omega} h(\omega).$$

But, the three sums here are precisely $\mathbf{P}(A)$, $\mathbf{P}(B)$ and $\mathbf{P}(A \cup B)$. Thus, we get $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

(3) This is similar to finite additivity, but needs a more involved argument. We leave it as an exercise for the interested reader. ■

Exercise 2. Adapt the proof to prove that for a countable family of events A_k in a common probability space (no disjointness assumed), we have

$$\mathbf{P}(\cup_k A_k) \leq \sum_k \mathbf{P}(A_k).$$

Definition 3 (Limsup and liminf of sets)**.** If $\{A_k, k \geq 1\}$, is a sequence of subsets of Ω, we define

$$\limsup A_k = \bigcap_{N=1}^{\infty} \bigcup_{k=N}^{\infty} A_k, \qquad \text{and} \qquad \liminf A_k = \bigcup_{N=1}^{\infty} \bigcap_{k=N}^{\infty} A_k.$$

In words, $\limsup A_k$ is the set of all ω that belong to infinitely many of the A_ks, and $\liminf A_k$ is the set of all ω that belong to all but finitely many of the A_ks.

Two special cases are of increasing and decreasing sequences of events. This means $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ and $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$. In these cases, the limsup and liminf are the same (so we refer to it as the limit of the sequence of sets). It is $\cup_k A_k$ in the case of increasing events and $\cap_k A_k$ in the case of decreasing events.

Exercise 4. (Monotonicity of $\mathbf{P}$) Events below are all contained in a discrete probability space. Use countable additivity of probability to show that

(1) If A_k are increasing events with limit A, show that $\mathbf{P}(A)$ is the increasing limit of $\mathbf{P}(A_k)$.

(2) If A_k are decreasing events with limit A, show that $\mathbf{P}(A)$ is the decreasing limit of $\mathbf{P}(A_k)$.

Now we re-write the basic rules of probability as follows:

The basic rules of probability:

(1) $\mathbf{P}(\emptyset) = 0$, $\mathbf{P}(\Omega) = 1$ and $0 \leq \mathbf{P}(A) \leq 1$ for any event A.

(2) $\mathbf{P}\left(\bigcup\limits_k A_k\right) \leq \sum\limits_k \mathbf{P}(A_k)$ for any countable collection of events A_k.

(3) $\mathbf{P}\left(\bigcup\limits_k A_k\right) = \sum\limits_k \mathbf{P}(A_k)$ if A_k is a countable collection of pairwise disjoint events.

2. INCLUSION-EXCLUSION FORMULA

In general, there is no simple rule for $\mathbf{P}(A \cup B)$ in terms of $\mathbf{P}(A)$ and $\mathbf{P}(B)$. Indeed, consider the probability space $\Omega = \{0,1\}$ with $p_0 = p_1 = \frac{1}{2}$. If $A = \{0\}$ and $B = \{1\}$, then $\mathbf{P}(A) = \mathbf{P}(B) = \frac{1}{2}$ and $\mathbf{P}(A \cup B) = 1$. However, if $A = B = \{0\}$, then $\mathbf{P}(A) = \mathbf{P}(B) = \frac{1}{2}$ as before, but $\mathbf{P}(A \cup B) = \frac{1}{2}$. This shows that $\mathbf{P}(A \cup B)$ cannot be determined from $\mathbf{P}(A)$ and $\mathbf{P}(B)$. Similarly for $\mathbf{P}(A \cap B)$ or other set constructions.

However, it is easy to see that $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$. This formula is not entirely useless, because in special situations we shall later see that the probability of the intersection is easy to compute and hence we may compute the probability of the union. Generalizing this idea to more than two sets, we get the following surprisingly useful formula.

Proposition 5 (**Inclusion-Exclusion formula**). *Let (Ω, p) be a probability space and let $A_1, \ldots, A_n$ be events. Then,*

$$\mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) = S_1 - S_2 + S_3 - \cdots + (-1)^{n-1} S_n,$$

where

$$S_k = \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \mathbf{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}).$$

We give two proofs, but the difference is only superficial. It is a good exercise to reason out why the two arguments are basically the same.

First proof. For each $\omega \in \Omega$, we compute its contribution to the two sides. If $\omega \notin \bigcup_{i=1}^{n} A_i$, then p_ω is not counted on either side. Suppose $\omega \in \bigcup_{i=1}^{n} A_i$ so that p_ω is counted once on the left side. We count the number of times p_ω is counted on the right side by splitting into cases depending on the exact number of A_is that contain ω.

Suppose ω belongs to exactly one of the A_is. For simplicity let us suppose that $\omega \in A_1$, but $\omega \in A_i^c$ for $2 \le i \le n$. Then p_ω is counted once in S_1 but not counted in $S_2, \ldots, S_n$.

Suppose ω belongs to A_1 and A_2 but not any other A_i. Then p_ω is counted twice in S_1 (once for $\mathbf{P}(A_1)$ and once for $\mathbf{P}(A_2)$) and subtracted once in S_2 (in $\mathbf{P}(A_1 \cap A_2)$). Thus, it is effectively counted once on the right side. The same holds if ω belongs to A_i and A_j but not any other A_ks.

If ω belongs to $A_1, \ldots, A_k$ but not any other A_i, then on the right side, p_ω is added k times in S_1, subtracted $\binom{k}{2}$ times in S_2, added $\binom{k}{3}$ times in S_k, and so on. Thus, p_ω is effectively counted

$$\binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \cdots + (-1)^{k-1}\binom{k}{k}$$

times. By the Binomial formula, this is just the expansion of $1 - (1-1)^k$ which is 1. ■

Second proof. Use the definition to write both sides of the statement. Let $A = \cup_{i=1}^n A_i$.

$$\text{LHS} = \sum_{\omega \in A} p_\omega = \sum_{\omega\in\Omega} \mathbf{1}_A(\omega) p_\omega.$$

Now, we compute the right side. For any $i_1 < i_2 < \cdots < i_k$, we write

$$\mathbf{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right) = \sum_{\omega\in\Omega} p_\omega \mathbf{1}_{A_{i_1}\cap\cdots\cap A_{i_k}}(\omega) = \sum_{\omega\in\Omega} p_\omega \prod_{\ell=1}^{k} \mathbf{1}_{A_{i_\ell}}(\omega).$$

Hence, the right hand side is given by adding over $i_1 < \cdots < i_k$, multiplying by $(-1)^{k-1}$ and then summing over k from 1 to n.

$$\begin{aligned}
\text{RHS} &= \sum_{k=1}^{n}(-1)^{k-1} \sum_{1\le i_1<\cdots<i_k\le n} \sum_{\omega\in\Omega} p_\omega \prod_{\ell=1}^{k} \mathbf{1}_{A_{i_\ell}}(\omega) \\
&= \sum_{\omega\in\Omega}\sum_{k=1}^{n}(-1)^{k-1} \sum_{1\le i_1<\cdots<i_k\le n} p_\omega \prod_{\ell=1}^{k} \mathbf{1}_{A_{i_\ell}}(\omega) \\
&= -\sum_{\omega\in\Omega} p_\omega \sum_{k=1}^{n} \sum_{1\le i_1<\cdots<i_k\le n} \prod_{\ell=1}^{k}(-\mathbf{1}_{A_{i_\ell}}(\omega)) \\
&= -\sum_{\omega\in\Omega} p_\omega \left(\prod_{j=1}^{n}(1-\mathbf{1}_{A_j}(\omega)) \; - \; 1\right) \\
&= \sum_{\omega\in\Omega} p_\omega \mathbf{1}_A(\omega)
\end{aligned}$$

because the quantity $\prod_{j=1}^{n}(1-\mathbf{1}_{A_j}(\omega))$ equals -1 if ω belongs to at least one of the A_is, and is zero otherwise. Thus the claim follows. ■

As we remarked earlier, it turns out that in many settings it is possible to compute the probabilities of intersections. We give an example now.

Example 6. Place n distinguishable balls in r distinguishable urns at random. Let A be the event that some urn is empty. The probability space is $\Omega = \{\underline{\omega} = (\omega_1, \ldots, \omega_n) : 1 \le \omega_i \le r\}$ with $p_{\underline{\omega}} = r^{-n}$. Let $A_\ell = \{\underline{\omega} : \omega_i \neq \ell\}$ for $\ell = 1, 2\ldots, r$. Then, $A = A_1 \cup \cdots \cup A_{r-1}$ (as A_r is empty, we could include it or not, makes no difference).

It is easy to see that $\mathbf{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = (r-k)^n r^{-n} = (1-\frac{k}{r})^n$. We could use the inclusion-exclusion formula to write the expression

$$\mathbf{P}(A) = r\left(1-\frac{1}{r}\right)^n - \binom{r}{2}\left(1-\frac{2}{r}\right)^n + \cdots + (-1)^{r-2}\binom{r}{r-1}\left(1-\frac{r-1}{r}\right)^n.$$

The last term is zero (since all urns cannot be empty). I donot know if this expression can be simplified any more.

We mention two useful formulas that can be proved on lines similar to the inclusion-exclusion principle. If we say "at least one of the events $A_1, A_2, \ldots, A_n$ occurs", we are talking about the union, $A_1 \cup A_2 \cup \cdots \cup A_n$. What about "at least m of the events $A_1, A_2, \ldots, A_n$ occur", how to express it with set operations. It is not hard to see that this set is precisely

$$B_m = \bigcup_{1 \le i_1 < i_2 < \cdots < i_m \le n} (A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}).$$

The event that "exactly m of the events $A_1, A_2, \ldots, A_n$ occur" can be written as

$$C_m = B_m \setminus B_{m+1} = \bigcup_{\substack{S \subseteq [n] \\ |S| = m}} \left(\bigcap_{i \in S} A_i \right) \bigcap \left(\bigcap_{i \notin S} A_i^c \right).$$

Exercise 7. Let $A_1, \ldots, A_n$ be events in a probability space (Ω, p) and let $m \le n$. Let B_m and C_m be as above. Show that

$$\begin{aligned}
\mathbf{P}(B_m) &= \sum_{k=m}^{n} (-1)^{k-m} \binom{k-1}{k-m} S_k \\
&= S_m - \binom{m}{1} S_{m+1} + \binom{m+1}{2} S_{m+2} - \binom{m+2}{3} S_{m+3} + \cdots \\
\mathbf{P}(C_m) &= \sum_{k=m}^{n} (-1)^{k-m} \binom{k}{m} S_k \\
&= S_m - \binom{m+1}{1} S_{m+1} + \binom{m+2}{2} S_{m+2} - \binom{m+3}{3} S_{m+3} + \cdots
\end{aligned}$$

3. Bonferroni's inequalities

Inclusion-exclusion formula is nice when we can calculate the probabilities of intersections of the events under consideration. Things are not always this nice, and sometimes that may be very difficult. Even if we could find them, summing them with signs according to the inclusion-exclusion formula may be difficult as Example 6 demonstrates. The *idea* behind the inclusion-exclusion formula can however be often used to compute *approximate values of probabilities*, which is very valuable in most applications. That is what we do next.

We know that $\mathbf{P}(A_1 \cup \cdots \cup A_n) \le \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n)$ for any events $A_1, \ldots, A_n$. This is an extremely useful inequality, often called the *union bound*. Its usefulness is in the fact that there is no assumption made about the events A_is (such as whether they are disjoint or not). The following inequalities generalize the union bound, and gives both upper and lower bounds for the probability of the union of a bunch of events.

Lemma 8 (Bonferroni's inequalities)**.** *Let $A_1, \ldots, A_n$ be events in a probability space (Ω, p) and let $A = A_1 \cup \cdots \cup A_n$. Then, $S_1 - S_2 \le \mathbf{P}(A) \le S_1$. More generally,*

$$\mathbf{P}(A) \le S_1 - S_2 + \cdots + S_m \quad \textit{if } m \textit{ is odd},$$

$$\mathbf{P}(A) \le S_1 - S_2 + \cdots - S_m \quad \textit{if } m \textit{ is even}.$$

Proof. We shall write out the proof for the cases $m = 1$ and $m = 2$. When $m = 1$, the inequality is just the union bound

$$\mathbf{P}(A) \le \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n)$$

which we know. When $m = 2$, the inequality to be proved is

$$\mathbf{P}(A) \ge \sum_k \mathbf{P}(A_k) - \sum_{k<\ell} \mathbf{P}(A_k \cap A_\ell)$$

To see this, fix $\omega \in \Omega$ and count the contribution of p_ω to both sides. Like in the proof of the inclusion-exclusion formula, for $\omega \notin A_1 \cup \cdots \cup A_n$, the contribution to both sides is zero. On the other hand, if ω belongs to exactly r of the sets for some $r \ge 1$, then it is counted once on the left side and $r - \binom{r}{2}$ times on the right side. Note that $r - \binom{r}{2} = \frac{1}{2}r(3 - r)$ which is always non-positive (one if $r = 1$, zero if $r = 2$ and non-positive if $r \ge 3$). Hence, we get LHS $\ge$ RHS.

Similarly, one can prove the other inequalities in the series. We leave it as an exercise. The key point is that $r - \binom{r}{2} + \cdots + (-1)^{k-1}\binom{r}{k}$ is non-negative if k is odd and non-positive if k is even (prove this). Here, as always, $\binom{x}{y}$ is interpreted as zero if $y > x$. ■

Here is an application of these inequalities.

Example 9. Return to Example 6. We obtained an exact expression for the answer, but that is rather complicated. For example, what is the probability of having at least one empty urn when

$n = 40$ balls are placed at random in $r = 10$ urns? It would be complicated to sum the series. Instead, we could use Bonferroni's inequalities to get the following bounds.

$$r\left(1-\frac{1}{r}\right)^n - \binom{r}{2}\left(1-\frac{2}{r}\right)^n \le \mathbf{P}(A) \le r\left(1-\frac{1}{r}\right)^n.$$

If we take $n = 40$ and $r = 10$, the bounds we get are $0.1418 \le \mathbf{P}(A) \le 0.1478$. Thus, we get a pretty decent approximation to the probability. By experimenting with other numbers you can check that the approximations are good when n is large compared to r but not otherwise. Can you reason why?

4. Independence - a first look

We remarked in the context of inclusion-exclusion formulas that often the probabilities of intersections of events is easy to find, and then we can use them to find probabilities of unions, etc. In many contexts, this is related to one of the most important notions in probability.

Definition 10. Let A, B be events in a common probability space. We say that A and B are *independent* is $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.

Caution: Independence should not be confused with disjointness! If A and B are disjoint, $\mathbf{P}(A \cap B) = 0$ and hence A and B can be independent if and only if one of $\mathbf{P}(A)$ or $\mathbf{P}(B)$ equals 0. Intuitively, if A and B are disjoint, then knowing that A occurred gives us a lot of information about B (that it did not occur!), so independence is not to be expected.

Example 11. Toss a fair coin n times. Then $\Omega = \{\underline{\omega} : \underline{\omega} = (\omega_1, \ldots, \omega_n),\ \omega_i \text{ is } 0 \text{ or } 1\}$ and $p_{\underline{\omega}} = 2^{-n}$ for each $\underline{\omega}$. Let $A = \{\underline{\omega} : \omega_1 = 0\}$ and let $B = \{\underline{\omega} : \omega_2 = 0\}$. Then, from the definition of probabilities, we can see that $\mathbf{P}(A) = 1/2$, $\mathbf{P}(B) = 1/2$ (because the elementary probabilities are equal, and both the sets A and B contain exactly 2^{n-1} elements). Further, $A \cap B = \{\underline{\omega} : \omega_1 = 1, \omega_2 = 0\}$ has 2^{n-2} elements, whence $\mathbf{P}(A \cap B) = 1/4$. Thus, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$ and hence A and B are independent.

If two events are independent, then the probability of their intersection can be found from the individual probabilities. How do we check if two events are independent? By checking if the probability of the event is equal to the product of the individual probabilities! It seems totally circular and useless! There are many reasons why it is not an empty notion as we shall see.

Firstly, in physical situations dependence is related to a basic intuition we have about whether two events are related or not. For example, suppose you are thinking of betting Rs.1000 on a particular horse in a race. If you get the news that your cousin is getting married, it will perhaps not affect the amount you plan to bet. However, if you get the news that one of the other horses has been injected with undetectable drugs, it might affect the bet you want to place. In other words, certain events (like marriage of a cousin) have no bearing on the probability of the event of interest (the event that our horse wins) while other events (like the injection of drugs) do have an impact. This intuition is often put into the very definition of probability space that we have.

For example, in the above example of tossing a fair coin n times, it is our intuition that a coin does not remember how it fell previous times, and that chance of its falling head in any toss is just $1/2$, irrespective of how many heads or tails occured before[1] And this intuition was used in

[1] It may be better to attribute this to experience rather than intuition. There have been reasonable people in history who believed that if a coin shows heads in ten tosses in a row, then on the next toss it is more likely to show tails (to 'compensate' for the overabundance of heads)! Clearly this is also someone's intuition, and different from ours. Only experiment can decide which is correct, and any number of experiments with real coins show that our intuition is correct, and coins have no memory.

defining the elementary probabilities as 2^{-n} each. Since we started with the intuitive notion of independence, and put that into the definition of the probability space, it is quite expected that the event that the first toss is a head should be independent of the event that the second toss is a tail. That is the calculation shown above.

But, how is independence useful mathematically if the conditions to check independence are the very conclusions we want?! The answer to this lies in the following fact (to be explained later). When certain events are independent, then many other collections of events that can be made out of them also turn out to be independent. For example, if A, B, C, D are independent (we have not yet defined what this means!), then $A \cup B$ and $C \cup D$ are also independent. Thus, starting from independence of certain events, we get independence of many other events. For example, any event depending on the first four tosses is independent of eny event depending on the next five tosses.

1. Conditional probability and independence

Definition 1. Let A, B be two events in the same probability space.

(1) If $\mathbf{P}(B) > 0$, we define the *conditional probability of A given B* as

$$\mathbf{P}\left(A \,\middle|\, B\right) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

(2) We say that A and B are *independent* if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. If $\mathbf{P}(B) \neq 0$, then A and B are independent if and only if $\mathbf{P}(A \mid B) = \mathbf{P}(A)$ (and similarly with the roles of A and B reversed). If $\mathbf{P}(B) = 0$, then A and B are necessarily independent since $\mathbf{P}(A \cap B)$ must also be 0.

What do these notions mean intuitively? In real life, we keep updating probabilities based on information that we get. For example, when playing cards, the chance that a randomly chosen card is an ace is $1/13$, but having drawn a card, the probability for the next card may not be the same - if the first card was seen to be an ace, then the chance of the second being an ace falls to $3/51$. This updated probability is called a conditional probability. Independence of two events A and B means that knowing whether or not A occurred does not change the chance of occurrence of B. In other words, the conditional probability of A given B is the same as the unconditional (original) probability of A.

Example 2. Let $\Omega = \{(i,j) : 1 \leq i, j \leq 6\}$ with $p_{(i,j)} = \frac{1}{36}$. This is the probability space corresponding to a throw of two fair dice. Let $A = \{(i,j) : i \text{ is odd}\}$ and $B = \{(i,j) : j \text{ is } 1 \text{ or } 6\}$ and $C = \{(i,j) : i + j = 4\}$. Then, $A \cap B = \{(i,j) : i = 1, 3, \text{ or } 5, \text{ and } j = 1 \text{ or } 6\}$. It is easy to see that

$$\mathbf{P}(A \cap B) = \frac{6}{36} = \frac{1}{6}, \quad \mathbf{P}(A) = \frac{18}{36} = \frac{1}{2}, \quad \mathbf{P}(B) = \frac{12}{36} = \frac{1}{3}.$$

In this case, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$ and hence A and B are independent. On the other hand,

$$\mathbf{P}(A \cap C) = \mathbf{P}\{(1,3),(3,1)\} = \frac{1}{18}, \quad \mathbf{P}(C) = \mathbf{P}\{(1,3),(2,2),(3,1)\} = \frac{1}{12}.$$

Thus, $\mathbf{P}(A \cap C) \neq \mathbf{P}(A)\mathbf{P}(C)$ and hence A and C are not independent.

This agrees with the intuitive understanding of independence since A is an event that depends only on the first toss and B is an event that depends only on the second toss. Therefore, A and B ought to be independent. However, C depends on both tosses, and hence cannot be expected to be independent of A. Indeed, it is easy to see that $\mathbf{P}(C \mid A) = \frac{1}{9}$.

Caution: Independence should not be confused with disjointness! If A and B are disjoint, $\mathbf{P}(A \cap B) = 0$ and hence A and B can be independent if and only if one of $\mathbf{P}(A)$ or $\mathbf{P}(B)$ equals 0. Intuitively, if A and B are disjoint, then knowing that A occurred gives us a lot of information about B (that it did not occur!), so independence is not to be expected.

Exercise 3. If A and B are independent events, show that the following pairs of events are also independent.

(1) A and B^c.

(2) A^c and B.

(3) A^c and B^c.

Total probability rule and Bayes' rule: Let $A_1, \dots, A_n$ be pairwise disjoint and mutually exhaustive events in a probability space. Assume $\mathbf{P}(A_i) > 0$ for all i. This means that $A_i \cap A_j = \emptyset$ for any $i \neq j$ and $A_1 \cup A_2 \cup \dots \cup A_n = \Omega$. We also refer to such a collection of events as a partition of the sample space.

Let B be any other event.

(1) (Total probability rule). $\mathbf{P}(B) = \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \dots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)$.

(2) (Bayes' rule). Assume that $\mathbf{P}(B) > 0$. Then, for each $k = 1, 2, \dots, n$, we have

$$\mathbf{P}(A_k \mid B) = \frac{\mathbf{P}(A_k)\mathbf{P}(B \mid A_k)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \dots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)}.$$

Proof. The proof is merely by following the definition.

(1) The right hand side is equal to

$$\mathbf{P}(A_1)\frac{\mathbf{P}(B \cap A_1)}{\mathbf{P}(A_1)} + \dots + \mathbf{P}(A_n)\frac{\mathbf{P}(B \cap A_n)}{\mathbf{P}(A_n)} = \mathbf{P}(B \cap A_1) + \dots + \mathbf{P}(B \cap A_n),$$

which is equal to $\mathbf{P}(B)$ since A_i are pairwise disjoint and exhaustive.

(2) Without loss of generality take $k = 1$. Note that $\mathbf{P}(A_1 \cap B) = \mathbf{P}(B \cap A_1) = \mathbf{P}(A_1)\mathbf{P}(B \mid A_1)$. Hence,

$$\begin{aligned}\mathbf{P}(A_1 \mid B) &= \frac{\mathbf{P}(A_1 \cap B)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \dots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)},\end{aligned}$$

where we used the total probability rule to get the denominator. ■

Exercise 4. Suppose A_i are events such that $\mathbf{P}(A_1 \cap \dots \cap A_n) > 0$. Then, show that

$$\mathbf{P}(A_1 \cap \dots \cap A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2 \mid A_1)\mathbf{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbf{P}(A_n \mid A_1 \cap \dots \cap A_{n-1}).$$

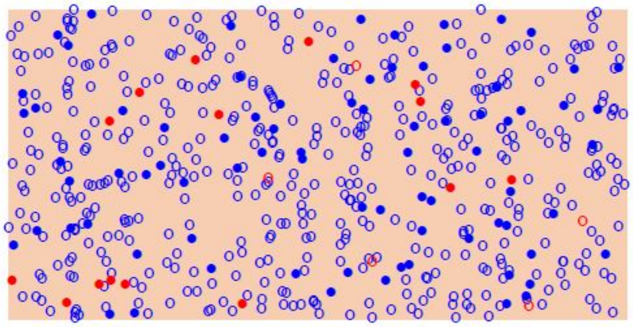


FIGURE 1. A population of healthy (blue) and diseased (red) individuals. Filled circle indicates those who tested positive and hollow circles indicate those who tested negative. The majority of those who tested positive are in fact healthy.

Example 5. Consider a rare disease X that affects one in a million people. A medical test is used to test for the presence of the disease. The test is 99% accurate in the sense that if a person has no disease, the chance that the test shows positive is 1% and if the person has disease, the chance that the test shows negative is also 1%.

Suppose a person is tested for the disease and the test result is positive. What is the chance that the person has the disease X?

Let A be the event that the person has the disease X. Let B be the event that the test shows positive. The given data may be summarized as follows.

(1) $\mathbf{P}(A) = 10^{-6}$. Of course $\mathbf{P}(A^c) = 1 - 10^{-6}$.

(2) $\mathbf{P}(B \mid A) = 0.99$ and $\mathbf{P}(B \mid A^c) = 0.01$.

What we want to find is $\mathbf{P}(A \mid B)$. By Bayes' rule (the relevant partition is $A_1 = A$ and $A_2 = A^c$),

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(B \mid A)\mathbf{P}(A)}{\mathbf{P}(B \mid A)\mathbf{P}(A) + \mathbf{P}(B \mid A^c)\mathbf{P}(A^c)} = \frac{0.99 \times 10^{-6}}{0.99 \times 10^{-6} + 0.01 \times (1 - 10^{-6})} = 0.000099.$$

The test is quite an accurate one, but the person tested positive has a really low chance of actually having the disease! Of course, one should observe that the chance of having disease is now approximately 10^{-4} which is considerably higher than 10^{-6}.

A calculation-free understanding of this surprising looking phenomenon can be achieved as follows: Let everyone in the population undergo the test. If there are 10^9 people in the population, then there are only 10^3 people with the disease. The number of true positives is approximately $10^3 \times 0.99 \approx 10^3$ while the number of false positives is $(10^9 - 10^3) \times 0.01 \approx 10^7$. In other words, among all positives, the false positives are way more numerous than true positives.

The surprise here comes from not taking into account the relative sizes of the sub-populations with and without the disease. Here is another manifestation of exactly the same fallacious reasoning.

Question: A person X is introverted, very systematic in thinking and somewhat absent-minded. You are told that he is a doctor or a mathematician. What would be your guess - doctor or mathematician?

Most people answer "mathematician". Even accepting the stereotype that a mathematician is more likely to have all these qualities than a doctor, this answer ignores the fact that there are perhaps a hundred times more doctors in the world than mathematicians! In fact, the situation is identical to the one in the example above, and the mistake is in confusing $\mathbf{P}(A\big|B)$ and $\mathbf{P}(B\big|A)$.

Medical diagnosis: Several different physiological problems can give rise to the same symptoms in a person. When a patient goes to a doctor and tells his/her symptoms, the doctor tries to guess the underlying disease that is causing the symptoms. This is Bayes' rule at work (or ought to be at work). Even though one may not be to write down all the probabilities, there is a lesson from the previous examples, which is that a priori chances of different diseases must be taken into account. In other words, suppose a rare but serious lung problem P always causes some symptom X to show. Suppose that common cold Q (rather common) also causes the symptom X in 1% of the cases.

If you are a doctor and you encounter a patient with symptom X, what would be your first guess - that it is caused by P, or by Q? Is such reasoning really used by doctors? It has been observed from various experiments that people do not naturally/intuitively do the right reasoning in such cases - they tend to overestimate the chance of the cause being the rare disease P.

2. Independence of three or more events

Definition 6. Events $A_1, \ldots, A_n$ in a common probability space are said to be independent if

$$\mathbf{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}\right) = \mathbf{P}(A_{i_1})\mathbf{P}(A_{i_2}) \cdots \mathbf{P}(A_{i_m})$$

for every choice of $m \le n$ and every choice of $1 \le i_1 < i_2 < \ldots < i_m \le n$.

The independence of n events requires us to check 2^n equations (that many choices of $i_1, i_2, \ldots$). Should it not suffice to check that each pair of A_i and A_j are independent? The following example shows that this is not the case!

Example 7. Let $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = 2^{-n}$ for each $\underline{\omega} \in \Omega$. Define the events $A = \{\underline{\omega} : \omega_1 = 0\}$, $B = \{\underline{\omega} : \omega_2 = 0\}$ and $C = \{\underline{\omega} : \omega_1 + \omega_2 = 0 \text{ or } 2\}$. In words, we toss a fair coin n times and A denotes the event that the first toss is a tail, B denotes the event that the second toss is a tail and C denotes the event that out of the first two tosses are both heads or both tails. Then $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(C) = \frac{1}{4}$. Further,

$$\mathbf{P}(A \cap B) = \frac{1}{4},\ \mathbf{P}(B \cap C) = \frac{1}{4},\ P(A \cap C) = \frac{1}{4},\ \mathbf{P}(A \cap B \cap C) = \frac{1}{4}.$$

Thus, A, B, C are independent *pairwise*, but not independent by our definition because $\mathbf{P}(A \cap B \cap C) \neq \frac{1}{8} = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$.

Intuitively this is right. Knowing A does not give any information about C (similarly with A and B, or B and C), but knowing A and B tells us completely whether or not C occurred! Thus it is right that the definition should not declare them to be independent.

Exercise 8. Let $A_1, \ldots, A_n$ be events in a common probability space. Then, $A_1, A_2, \ldots, A_n$ are independent if and only if the following equalities hold: For each i, define B_i as A_i or A_i^c. Then

$$\mathbf{P}(B_1 \cap B_2 \cap \cdots \cap B_n) = \mathbf{P}(B_1)\mathbf{P}(B_2) \cdots \mathbf{P}(B_n).$$

Note: This should hold for any possible choice of B_is. In other words, the system of 2^n equalities in the definition of independence may be replaced by this new set of 2^n equalities. The latter system has the advantage that it immediately tells us that if $A_1, \ldots, A_n$ are independent, then $A_1, A_2^c, A_3, \ldots$ (for each i choose A_i, or its complement) are independent.

3. Discrete probability distributions

Let (Ω, p) be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable. We define two objects associated to X.

Probability mass function (pmf). The range of X is a countable subset of $\mathbb{R}$, denote it by $\text{Range}(X) = \{t_1, t_2, \ldots\}$. Then, define $f_X : \mathbb{R} \to [0,1]$ as the function

$$f_X(t) = \begin{cases} \mathbf{P}\{\omega : X(\omega) = t\} & \text{if } t \in \text{Range}(X), \\ 0 & \text{if } t \notin \text{Range}(X). \end{cases}$$

One obvious property is that $\sum_{t \in \mathbb{R}} f_X(t) = 1$. Conversely, any non-negative function f that is non-zero on a countable set S and such that $\sum_{t \in \mathbb{R}} f(t) = 1$ is a pmf of some random variable.

Cumulative distribution function (CDF). Define $F_X : \mathbb{R} \to [0,1]$ by

$$F_X(t) = \mathbf{P}\{\omega : X(\omega) \le t\} \text{ for } t \in \mathbb{R}.$$

Example 9. Let $\Omega = \{(i,j) : 1 \le i, j \le 6\}$ with $p_{(i,j)} = \frac{1}{36}$ for all $(i,j) \in \Omega$. Let $X : \Omega \to \mathbb{R}$ be the random variable defined by $X(i,j) = i + j$. Then, $\text{Range}(X) = \{2, 3, \ldots, 12\}$. The pmf of X is given by

k	2	3	4	5	6	7	8	9	10	11	12
$f_X(k)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

and the CDF is given by

t	< 2	$[2,3)$	$[3,4)$	$[4,5)$	$[5,6)$	$[6,7)$	$[7,8)$	$[8,9)$	$[9,10)$	$[10,11)$	$[11,12)$	≥ 12
$F_X(t)$	0	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	1

Basic properties of a CDF: The following observations are easy to make:

(1) F is an increasing function on $\mathbb{R}$.

(2) $\lim\limits_{t \to +\infty} F(t) = 1$ and $\lim\limits_{t \to -\infty} F(t) = 0$.

(3) F is right continuous, that is, $\lim\limits_{h \to 0^+} F(t+h) = F(t+) = F(t)$ for all $t \in \mathbb{R}$.

(4) F increases only in jumps. This means that if F has no jump discontinuities (an increasing function has no other kind of discontinuity anyway) in an interval $[a,b]$, then $F(a) = F(b)$.

Since $F(t)$ is the probability of a certain event, these statements can be proved using the basic rules of probability that we saw earlier.

Proof. Let $t < s$. Define two events, $A = \{\omega : X(\omega) \le t\}$ and $B = \{\omega : X(\omega) \le s\}$. Clearly, $A \subseteq B$ and hence $F(t) = \mathbf{P}(A) \le \mathbf{P}(B) = F(s)$. This proves the first property.

To prove the second property, let $A_n = \{\omega : X(\omega) \le n\}$ for $n \ge 1$. Then, A_n are increasing in n and $\bigcup_{n=1}^{\infty} A_n = \Omega$. Hence, $F(n) = \mathbf{P}(A_n) \to \mathbf{P}(\Omega) = 1$ as $n \to \infty$. Since F is increasing, it follows that $\lim_{t\to+\infty} F(t) = 1$. Similarly, one can prove that $\lim_{t\to-\infty} F(t) = 0$.

Right continuity of F is also proved the same way, by considering the events $B_n = \{\omega : X(\omega) \le t + \frac{1}{n}\}$. We omit details. ■

Remark 10. It is easy to see that one can recover the pmf from the CDF and vice versa. For example, given the pmf f, we can write the CDF as $F(t) = \sum_{u:u\le t} f(u)$. Conversely, given the CDF, by looking at the locations of the jumps and the sizes of the jumps, we can recover the pmf.

The point is that *probabilistic questions about X can be answered by knowing its CDF F_X*. Therefore, in a sense, the probability space becomes irrelevant. For example, the expected value of a random variable can be computed using its CDF only. Hence, we shall often make statements like "X is a random variable with pmf f" or "X is a random variable with CDF F", without bothering to indicate the probability space.

Some distributions (i.e., CDF or the associated pmf) occur frequently enough to merit a name.

Example 11. Let f and F be the pmf, CDF pair

$$f_X(t) = \begin{cases} p & \text{if } t = 1, \\ q & \text{if } t = 0, \end{cases} \qquad F_X(t) = \begin{cases} 1 & \text{if } t \ge 1, \\ q & \text{if } t \in [0,1), \\ 0 & \text{if } t < 0. \end{cases}$$

A random variable X having this pmf (or, equivalently the CDF) is said to have *Bernoulli distribution* with parameter p (with $q = 1 - p$) and write $X \sim \text{Ber}(p)$. For example, if $\Omega = \{1, 2, \dots, 10\}$ with $p_i = 1/10$, and $X(\omega) = \mathbf{1}_{\omega\le 3}$, then $X \sim \text{Ber}(0.3)$. Any random variable taking only the values 0 and 1, has Bernoulli distribution.

Example 12. Fix $n \ge 1$ and $p \in [0,1]$. The pmf defined by $f(k) = \binom{n}{k} p^k q^{n-k}$ for $0 \le k \le n$ is called the *Binomial distribution* with parameters n and p, and is denoted Bin(n,p). The CDF is as usual defined by $F(t) = \sum_{\mathbf{u}:\mathbf{u}\le t} f(u)$, but it does not have any particularly nice expression.

For example, if $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = p^{\sum_i \omega_i} q^{n-\sum_i \omega_i}$ and $X(\underline{\omega}) = \omega_1 + \cdots + \omega_n$, then $X \sim \text{Bin}(n,p)$. In words, the number of heads in n tosses of a p-coin has Bin(n,p) distribution.

Example 13. Fix $p \in (0,1]$ and let $f(k) = q^{k-1}p$ for $k \in \mathbb{N}_+$. This is called the *Geometric distribution* with parameter p and is denoted Geo(p). The CDF is

$$F(t) = \begin{cases} 0 & \text{if } t < 1, \\ 1 - q^k & \text{if } k \le t < k+1, \text{ for some } k \ge 1. \end{cases}$$

For example, the number of tosses of a p-coin till the first head turns up, is a random variable with Geo(p) distribution.

Example 14. Fix $\lambda > 0$ and define the pmf $f(k) = e^{-\lambda}\frac{\lambda^k}{k!}$. This is called the *Poisson distribution* with parameter λ and is denoted Pois(λ).

Example 15. Fix positive integers b, w and $m \leq b + w$. Define the pmf $f(k) = \frac{\binom{b}{k}\binom{w}{m-k}}{\binom{b+w}{m}}$ where the binomial coefficient $\binom{x}{y}$ is interpreted to be zero if $y > x$ (thus $f(k) > 0$ only for $\max\{m - w, 0\} \leq k \leq b$). This is called the *Hypergeometric distribution* with parameters b, w, m and we shall denote it by Hypergeo(b, w, m).
Consider a population with b men and w women. The number of men in a random sample (without replacement) of size m, is a random variable with the Hypergeo(b, w, m) distribution.

4. General probability distributions

We take the first three of the four properties of CDF proved in the previous section as the *definition* of a CDF or distribution function, in general.

Definition 16. A (cumulative) distribution function (or, CDF for short) is any function $F : \mathbb{R} \to [0,1]$ be a non-decreasing, right continuous function such that $F(t) \to 0$ as $t \to -\infty$ and $F(t) \to 1$ as $t \to +\infty$.

If (Ω, p) is a discrete probability space and $X : \Omega \mapsto \mathbb{R}$ is any random variable, then the function $F(t) = \mathbf{P}\{\omega : X(\omega) \leq t\}$ is a CDF, as discussed in the previous section. However, there are distribution functions that do not arise in this manner.

Example 17. Let

$$F(t) = \begin{cases} 0 & \text{if } t \leq 0, \\ t & \text{if } 0 < t < 1, \\ 1 & \text{if } t \geq 1. \end{cases}$$

Then, it is easy to see that F is a distribution function. However, it has no jumps and hence it does not arise as the CDF of any random variable on a discrete probability space.

There are two ways to rectify this issue:

(1) The first way is to learn the notion of uncountable probability spaces, which poses many subtleties. It requires a semester or so of real analysis and measure theory. But, after that one can define random variables on uncountable probability spaces and the above example will turn out to be the CDF of some random variable on some (uncountable) probability space.

(2) Just regard CDFs such as in the above example as reasonable approximations to CDFs of some discrete random variables. For example, if $\Omega = \{\omega_0, \omega_1, \ldots, \omega_N\}$ and $p(\omega_k) = 1/(N+1)$ for all $0 \leq k \leq N$, and $X : \Omega \mapsto \mathbb{R}$ is defined by $X(\omega_k) = k/N$, then it is easy to check that the CDF of X is the function G given by

$$G(t) = \begin{cases} 0 & \text{if } t \leq 0, \\ \frac{k}{N+1} & \text{if } \frac{k-1}{N} \leq t < \frac{k}{N} \text{ for some } k = 1, 2, \ldots, N, \\ 1 & \text{if } t \geq 1. \end{cases}$$

Now, if N is very large, then the function G looks approximately like the function F. Just as it is convenient to regard water as a continuous medium in some problems (although water is made up of molecules and is discrete at small scales), it is convenient to use the continuous function F as a reasonable approximation to the step function G.

We shall take the second option. Whenever we write continuous distribution functions such as in the above example, at the back of our mind we have a discrete random variable (taking a large number of closely placed values) whose CDF is approximated by our distribution function. The advantage of using continuous objects instead of discrete ones is that the powerful tools of Calculus become available to us.

5. Uncountable probability spaces - conceptual difficulties

The following two "random experiments" are easy to imagine, but difficult to fit into the framework of probability spaces.

(1) Toss a p-coin infinitely many times: Clearly the sample space is $\Omega = \{0,1\}^{\mathbb{N}}$. But what is $p_{\underline{\omega}}$ for any $\underline{\omega} \in \Omega$? The only reasonable answer is $p_{\underline{\omega}} = 0$ for all ω. But then how to define $\mathbf{P}(A)$ for any A? For example, if $A = \{\underline{\omega} : \omega_1 = 0, \omega_2 = 0, \omega_3 = 1\}$, then everyone agrees that $\mathbf{P}(A)$ "ought to be" q^2p, but how does that come about? The basic problem is that Ω is uncountable, and probabilities of events cannot be obtained by summing probabilities of singletons.

(2) Draw a number at random from $[0,1]$: Again, it is clear that $\Omega = [0,1]$, but it also seems reasonable that $p_x = 0$ for all x. Again, Ω is uncountable, and probabilities of events cannot be obtained by summing probabilities of singletons. It is "clear" that if $A = [0.1, 0.4]$, then $\mathbf{P}(A)$ "ought to be" 0.3, but it gets confusing when one tries to derive this from something more basic!

The resolution: Let Ω be uncountable. There is a class of *basic subsets* (usually NOT singletons) of Ω for which we take the probabilities as given. We also take the rules of probability, namely, countable additivity, as axioms. Then, we use the rules to compute the probabilities of more complex events (subsets of Ω) by expressing those events in terms of the basic sets using countable intersections, unions and complements and applying the rules of probability.

Example 18. In the example of infinite sequence of tosses, $\Omega = \{0,1\}^{\mathbb{N}}$. Any set of the form $A = \{\underline{\omega} : \omega_1 = \epsilon_1, \dots, \omega_k = \epsilon_k\}$, where $k \geq 1$ and $\epsilon_i \in \{0,1\}$ will be called a basic set and its probability is defined to be $\mathbf{P}(A) = \prod_{j=1}^{k} p^{\epsilon_j} q^{1-\epsilon_j}$, where we assume that $p > 0$. Now, consider a more complex event, for example, $B = \{\underline{\omega} : \omega_k = 1 \text{ for some } k\}$. We can write $B = A_1 \cup A_2 \cup A_3 \cup \cdots$, where $A_k = \{\underline{\omega} : \omega_1 = 0, \dots, \omega_{k-1} = 0, \omega_k = 1\}$. Since A_k are pairwise disjoint, the rules of probability demand that $\mathbf{P}(B)$ should be $\sum_k \mathbf{P}(A_k) = \sum_k q^{k-1}p$ which is in fact equal to 1.

Example 19. In the example of drawing a number at random from $[0,1]$, $\Omega = [0,1]$. Any interval (a,b) with $0 \leq a < b \leq 1$ is called a basic set and its probability is defined as $\mathbf{P}(a,b) = b - a$. Now, consider a non-basic event $B = [a,b]$. We can write $B = A_1 \cup A_2 \cup A_3 \dots$, where $A_k = (a + (1/k), b - (1/k))$. Then A_k is an increasing sequence of events and the rules of probability say that $\mathbf{P}(B)$ must be equal to $\lim_{k\to\infty} \mathbf{P}(A_k) = \lim_{k\to\infty}(b - a - (2/k)) = b - a$. Another example could be $C = [0.1, 0.2) \cup (0.3, 0.7]$. Similarly, argue that $\mathbf{P}(\{x\}) = 0$ for any $x \in [0,1]$. A more interesting one is $D = \mathbb{Q} \cap [0,1]$. Since it is a countable union of singletons, it must have zero probability! Even more interesting is the $1/3$-Cantor set. Although uncountable, it has zero probability! We discuss an example (related to the Cantor set) later.

Consistency: Is this truly a solution to the question of uncountable spaces? Are we assured of never running into inconsistencies? NOT always!

Example 20. Let $\Omega = [0,1]$ and let intervals (a,b) be open sets with their probabilities defined as $\mathbf{P}(a,b) = \sqrt{b-a}$. This quickly leads to problems. For example, $\mathbf{P}(0,1) = 1$ by definition. But $(0,1) = (0,0.5) \cup (0.5,1) \cup \{1/2\}$ from which the rules of probability would imply that $\mathbf{P}(0,1)$ must be at least $\mathbf{P}(0,1/2) + \mathbf{P}(1/2,1) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$ which is greater than 1. Inconsistency!

Exercise 21. Show that we run into inconsistencies if we define $\mathbf{P}(a,b) = (b-a)^2$ for $0 \leq a < b \leq 1$.

Subtleties of conditional probability

Conditional probabilities are quite subtle. Apart from the common mistake of confusing $\mathbf{P}(A \bigm| B)$ for $\mathbf{P}(B \bigm| A)$, there are other points one sometimes overlooks. In fact, most of the paradoxical sounding puzzles in probability are based on confusing aspects of conditional probability. Let us see one.

Question 9. A man says "I have two children, and one of them is a boy". What is the chance that the other one is a girl?

There are four possibilities BB, BG, GB, GG, of which GG has been eliminated. Of the remaining three, two are favourable, hence the chance is $2/3$ that the other child is a girl. This is a possible solution. If you accept this as reasonable, here is another question.

Question 10. A man says "I have two children, and one of them is a boy born on a Monday". What is the chance that the other one is a girl?

Does the addition of the information about the boy change the probability? One opinion is that it should not. The other is to follow the same solution pattern as before. Write down all the $2\times 2\times 7\times 7$ possibilities: $BBss$ (boy, boy, Sunday, Sunday), $BBsm$ (boy, boy, Sunday, Monday), etc. The given information that one is a boy who was born on Sunday eliminates many possibilities and what remain are 27 possibilities $BGm*$, $GB * m$, $BBm*$, $BB * m$ where $*$ is any day of the week. Take care to not double count $BBmm$ to see that there are 27 possibilities. Of these, 14 are favourable (i.e., the other child is a girl), hence we conclude that the probability is $14/27$.

Is the correct answer $14/27$ as calculated here or is it $2/3$? Since the information of the day of birth of the boy is irrelevant, why should we change our earlier answer of $2/3$?

Both answers can be correct, depending on the interpretation of what the experiment is. The main reason for all the confusion leading to multiple interpretations is avoided if one realizes this: *To compute conditional probabilities, it is not enough to know what the person said, but also what else he could have said*. Not realizing this point is the main source of confusion in many popular puzzles in probability. We leave you to understand this statement in the context of the problem, but explain it in more general terms.

What are we conditioning on? In talking about conditional probability, we should really think of a *measurement*. To start, we have a discrete probability space (Ω, p) which defines probabilities of various events. A measurement is a function $T : \Omega \mapsto \mathbb{R}$. Let us say that the measurement can take three values, $0, 1, 2$. Let $A_i = \{\omega : T(\omega) = i\}$, for $i = 1, 2, 3$. Then, A_1, A_2, A_3 are pairwise disjoint and their union is Ω. As a result of the measurement, we get to know whether ω belongs to A_1 or to A_2 or to A_3. But, we would not know which of these it belongs to. Based on what the measurement actually shows, we update our probabilities.

The problem with puzzles (like the one above) is that they don't specify what is being measured. Depending on the interpretation, different answers are possible. For example, if a person says "I have two children, one of whom is a girl", it is giving the result of a measurement without saying what was being measured. Did the person check the sex of the eldest child and report "girl" as the measurement? Did he check whether or not he has a girl child and then report "Yes" as the measurement?

One lesson from this is this. We should not think of $\mathbf{P}(\cdot \mid A)$ alone, but of both $\mathbf{P}(\cdot \mid A)$ and $\mathbf{P}(\cdot \mid A^c)$. We make a measurement which corners ω into A or into A^c, and we have to be ready to deploy $\mathbf{P}(\cdot \mid A)$ or $\mathbf{P}(\cdot \mid A^c)$, depending on the outcome of that measurement.

1. UNCOUNTABLE PROBABILITY SPACES - THE RESOLUTION

Recall the example of drawing a number at random from $\Omega = [0,1]$. Any interval (a,b) with $0 \leq a < b \leq 1$ is called a basic set and its probability is defined as $\mathbf{P}(a,b) = b - a$. Is this truly a solution to the question of uncountable spaces? Are we assured of never running into inconsistencies? NOT always!

Example 1. Let $\Omega = [0,1]$ and let intervals (a,b) be open sets with their probabilities defined as $\mathbf{P}(a,b) = \sqrt{b-a}$. This quickly leads to problems. For example, $\mathbf{P}(0,1) = 1$ by definition. But $(0,1) = (0,0.5) \cup (0.5,1) \cup \{1/2\}$ from which the rules of probability would imply that $\mathbf{P}(0,1)$ must be at least $\mathbf{P}(0,1/2) + \mathbf{P}(1/2,1) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$ which is greater than 1. Inconsistency!

Thus, one cannot arbitrarily assign probabilities to basic events. However, if we use the notion of distribution function to assign probabilities to intervals, then no inconsistencies arise.

Theorem 2. *Let $\Omega = \mathbb{R}$ and let intervals of the form $(a,b]$ with $a < b$ be called basic sets. Let F be any distribution function.* Define *the probabilities of basic sets as $\mathbf{P}\{(a,b]\} = F(b) - F(a)$. Then, applying the rules of probability to compute probabilities of more complex sets (obtained by taking countable intersections, unions and complements) will never lead to inconsistency.*

Let F be a CDF. Then, the above consistency theorem really asserts that there exists (a possibly uncountable) probability space and a random variable such that $F(t) = \mathbf{P}\{X \leq t\}$ for all t. We say that X has distribution F. However, it takes a lot of technicalities to define what uncountable probability spaces look like and what random variables mean in this more general setting, we shall never define them.

The job of a probabilist consists in taking a CDF F (then the probabilities of intervals are already given to us as $F(b) - F(a)$ etc.) and find probabilities of more general subsets of $\mathbb{R}$. Instead we can use the following simple working rules to answer questions about the distribution of a random variable. Here are the working rules:

(1) For an $a < b$, we set $\mathbf{P}\{a < X \leq b\} := F(b) - F(a)$.

(2) If $I_j = (a_j, b_j]$ are countably many pairwise disjoint intervals, and $I = \bigcup_j I_j$, then we define $\mathbf{P}\{X \in I\} := \sum_j F(b_j) - F(a_j)$.

(3) For a general set $A \subseteq \mathbb{R}$, here is a general scheme: Find countably many pairwise disjoint intervals $I_j = (a_j, b_j]$ such that $A \subseteq \cup_j I_j$. Then we define $\mathbf{P}\{X \in A\}$ as the infimum (over all such coverings by intervals) of the quantity $\sum_j F(b_j) - F(a_j)$.

All of probability in another line: Take an (interesting) random variable X with a given CDF F and an (interesting) set $A \subseteq \mathbb{R}$. Find $\mathbf{P}\{X \in A\}$.

There are loose threads here, but they can be safely ignored for this course. We just remark about them for those who are curious to know.

Remark 3. The above method starts from a CDF F and defines $\mathbf{P}\{X \in A\}$ for all subsets $A \subseteq \mathbb{R}$. However, for most choices of F, the countable additivity property turns out to be violated! The sets which violate them rarely arise in practice, and hence we ignore them for the present exposition.

Exercise 4. Let X be a random variable with distribution F. Use the working rules to find the following probabilities:

(1) Show that $\mathbf{P}\{X < a\} = F(a-)$, where $\lim\limits_{h \to 0^+} F(a-h) = F(a-)$.

Hint: $(-\infty, a - 1/n]$ increases to $(-\infty, a)$ as $n \to \infty$.

(2) Now, show that $\mathbf{P}\{X = a\} = F(a) - F(a-)$. In particular, this probability is zero unless F has a jump at a.

(3) Write $\mathbf{P}\{a < X < b\}$, $\mathbf{P}\{a \le X < b\}$, $\mathbf{P}\{a \le X \le b\}$ in terms of F.

We now illustrate how to calculate the probabilities of rather non-trivial sets in a special case. It is not always possible to get an explicit answer as here.

Example 5. Let F be the CDF defined below:

$$F(t) = \begin{cases} 0 & \text{if } t \le 0, \\ t & \text{if } 0 < t < 1, \\ 1 & \text{if } t \ge 1. \end{cases}$$

We calculate $\mathbf{P}\{X \in A\}$ for two sets A.

1. $A = \mathbb{Q} \cap [0,1]$. Since A is countable, we may write $A = \cup_n \{r_n\}$ and hence $A \subseteq \cup_n I_n$, where $I_n = (r_n, r_n + \delta 2^{-n}]$ for any fixed $\delta > 0$. Hence, $0 \le \mathbf{P}\{X \in A\} \le \sum_n F(r_n + \delta 2^{-n}) - F(r_n) \le 2\delta$. Since this is true for every $\delta > 0$, we must have $\mathbf{P}\{X \in A\} = 0$. (We stuck to the letter of the recipe described earlier. It would have been simpler to say that any countable set is a countable union of singletons, and by the countable additivity of probability, must have probability zero. Here, we have used the fact that singletons have zero probability since F is continuous).

2. $A =$ Cantor's set. [1] How to find $\mathbf{P}\{X \in A\}$? Let A_n be the set of all $x \in [0,1]$ which do not have 1 in the first n digits of their ternary expansion. Then $A \subseteq A_n$. Further, it is not hard to see that $A_n = I_1 \cup I_2 \cup \cdots \cup I_{2^n}$, where each of the intervals I_j has length equal to 3^{-n}. Therefore, $0 \leq \mathbf{P}\{X \in A\} \leq \mathbf{P}\{X \in A_n\} = 2^n 3^{-n}$ which goes to 0 as $n \to \infty$. Hence, $\mathbf{P}\{X \in A\} = 0$.

[1] To define the Cantor set, recall that any $x \in [0,1]$ may be written in ternary expansion as $x = 0.u_1 u_2 \ldots := \sum_{n=1}^{\infty} u_n 3^{-n}$ where $u_n \in \{0,1,2\}$. This expansion is unique except if x is a rational number of the form $p/3^m$ for some integers p, m (these are called triadic rationals). For triadic rationals, there are two possible ternary expansions, a terminating one and a non-terminating one (for example, $x = 1/3$ can be written as $0.100\ldots$ or as $0.0222\ldots$). For definiteness, for triadic rationals we shall always take the non-terminating ternary expansion. With this preparation, the Cantor set is defined as the set of all x which do not have the digit 1 in their ternary expansion.

2. Examples of continuous distributions

Cumulative distributions (CDF) will also be referred to as simply distribution functions (DF). We start by giving two large classes of CDFs. There are CDFs that do not belong to either of these classes, but for practical purposes they will be ignored (for now).

(1) (CDFs with pmf). Let f be a pmf, i.e., let $t_1, t_2, \dots$ be a countable subset of reals and let $f(t_i)$ be non-negative numbers such that $\sum_i f(t_i) = 1$. Define $F : \mathbb{R} \to [0,1]$ by

$$F(t) := \sum_{i : t_i \le t} f(t_i).$$

Then, F is a CDF. Indeed, we have seen that it is the CDF of a discrete random variable. A special feature of this CDF is that it increases only in jumps (in more precise language, if F is continuous on an interval $[s,t]$, then $F(s) = F(t)$).

(2) (CDFs with pdf). Let $f : \mathbb{R} \to \mathbb{R}_+$ be a function (convenient to assume that it is a piecewise continuous function) such that $\int_{-\infty}^{+\infty} f(u)du = 1$. Such a function is called a *probability density function* (or, pdf for short). Then, define $F : \mathbb{R} \to [0,1]$ by

$$F(t) := \int_{-\infty}^{t} f(u)du.$$

Again, F is a CDF. Indeed, it is clear that F has the increasing property (if $t > s$, then $F(t) - F(s) = \int_s^t f(u)du$ which is non-negative because $f(u)$ is non-negative for all u), and its limits at $\pm\infty$ are as they should be (Why?). As for right-continuity, F is in-fact continuous. Actually, F is differentiable except at points where f is discontinuous and $F'(t) = f(t)$.

Remark 6. We understand the pmf. For example if X has pmf f, then $f(t_i)$ is just the probability that X takes the value t_i. How to interpret the pdf? If X has pdf f, then as we already remarked, the CDF is continuous and hence $\mathbf{P}\{X = t\} = 0$. Therefore, $f(t)$ cannot be interpreted as $\mathbf{P}\{X = t\}$ (in fact, pdf can take values greater than 1, so it cannot be a probability!).

To interpret $f(a)$, take a small positive number δ and look at

$$F(a+\delta) - F(a) = \int\limits_{a}^{a+\delta} f(u)du \approx \delta f(a).$$

In other words, $f(a)$ measures the chance of the random variable taking values near a. Higher the pdf, greater the chance of taking values near that point.

Among distributions with pmf, recall that we have seen the Binomial, Poisson, Geometric and Hypergeometric families of distributions. Now, we give many important examples of distributions (CDFs) with densities.

Example 7. Uniform distribution on the interval $[a, b]$: Denoted by Unif($[a, b]$), where $a < b$ is the distribution with density and distribution given by

$$\text{PDF: } f(t) = \begin{cases} \frac{1}{b-a} & \text{if } t \in (a, b) \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \leq a \\ \frac{t-a}{b-a} & \text{if } t \in (a, b) \\ 1 & \text{if } t \geq b. \end{cases}$$

Example 8. Exponential distribution with parameter λ: Denoted by Exp(λ), where $\lambda > 0$ is the distribution with density and distribution given by

$$\text{PDF: } f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ 1 - e^{-\lambda t} & \text{if } t > 0. \end{cases}$$

Example 9. Normal distribution with parameters μ, σ^2: Denoted by N(μ, σ^2), where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ is the distribution with density and distribution given by

$$\text{PDF: } \varphi_{\mu,\sigma^2}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2} \qquad \text{CDF: } \Phi_{\mu,\sigma^2}(t) = \int\limits_{-\infty}^{t} \varphi_{\mu,\sigma^2}(u)du.$$

There is no closed form expression for the CDF. It is standard notation to write φ and Φ to denote the normal density and CDF when $\mu = 0$ and $\sigma^2 = 1$. $\mathrm{N}(0, 1)$ is called the standard normal distribution. By a change of variable, one can check that $\Phi_{\mu,\sigma^2}(t) = \Phi(\frac{t-\mu}{\sigma})$.

We said that the normal CDF has no simple expression, but is it even clear that it is a CDF?! In other words, is the proposed density a true pdf? Clearly $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ is non-negative. We need to check that its integral is 1.

Lemma 10. *Fix $\mu \in \mathbb{R}$ and $\sigma > 0$ and let $\varphi(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2}$. Then, $\int\limits_{-\infty}^{\infty} \varphi(t)dt = 1$.*

Proof. It suffices to check the case $\mu = 0$ and $\sigma^2 = 1$ (Why?). To find its integral is quite non-trivial. Let $I = \int_{-\infty}^{\infty} \varphi(t)dt$. We introduce the two-variable function $h(t, s) := \varphi(t)\varphi(s) = (2\pi)^{-1} e^{-(t^2+s^2)/2}$. On one hand,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(t, s)dtds = \left(\int_{-\infty}^{+\infty} \varphi(t)dt \right) \left(\int_{-\infty}^{+\infty} \varphi(s)ds \right) = I^2.$$

On the other hand, using polar co-ordinates $t = r\cos\theta$, $s = r\sin\theta$, we see that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(t, s)dtds = \int_{0}^{\infty} \int_{0}^{2\pi} (2\pi)^{-1} e^{-r^2/2} r d\theta dr = \int_{0}^{\infty} r e^{-r^2/2} dr = 1$$

since $\frac{d}{dr}e^{-r^2/2} = -re^{-r^2/2}$. Thus $I^2 = 1$, and hence $I = 1$. ■

Example 11. Gamma distribution with shape parameter ν and scale parameter λ: Denoted by Gamma(ν, λ) with $\nu > 0$ and $\lambda > 0$, is the distribution with density and distribution given by:

$$\text{PDF: } f(t) = \begin{cases} \frac{1}{\Gamma(\nu)}\lambda^\nu t^{\nu-1}e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ \int_0^t f(u)du & \text{if } t > 0. \end{cases}$$

Here, $\Gamma(\nu) := \int_0^\infty t^{\nu-1}e^{-t}dt$. Firstly, f is a density, i.e., it integrates to 1. To see this, make the change of variable $\lambda t = u$ to see that

$$\int_0^\infty \lambda^\nu e^{-\lambda t}t^{\nu-1}dt = \int_0^\infty e^{-u}u^{\nu-1}du = \Gamma(\nu).$$

Thus, $\int_0^\infty f(t)dt = 1$.

When $\nu = 1$, we get back the exponential distribution. Thus, the Gamma family subsumes the exponential distributions.

Exercise 12. For positive integer values of ν, one can actually write an expression for the CDF of Gamma(ν, λ) as

$$F_{\nu,\lambda}(t) = 1 - e^{-\lambda t}\sum_{k=0}^{\nu-1}\frac{(\lambda t)^k}{k!}.$$

Once the expression is given, it is easy to check it by induction (and integration by parts). A curious observation is that the right hand side is exactly $\mathbf{P}(N \geq \nu)$, where $N \sim$ Pois(λt). This is in fact indicating a deep connection between Poisson distribution and the Gamma distributions. The function $\Gamma(\nu)$, also known as Euler's Gamma function, is an interesting and important one and occurs all over mathematics. [2]

[2] **The Gamma function:** The function $\Gamma : (0,\infty) \to \mathbb{R}$ defined by $\Gamma(\nu) = \int_0^\infty e^{-t}t^{\nu-1}dt$ is a very important function that often occurs in mathematics and physics. There is no simpler expression for it, although one can find it explicitly for special values of ν. One of its most important properties is that $\Gamma(\nu+1) = \nu\Gamma(\nu)$. To see this, consider

$$\Gamma(\nu+1) = \int_0^\infty e^{-t}t^\nu dt = -e^{-t}t^\nu \Big|_0^\infty + \nu\int_0^\infty e^{-t}t^{\nu-1}dt = \nu\Gamma(\nu).$$

Starting with $\Gamma(1) = 1$ (direct computation) and using the above relationship repeatedly one sees that $\Gamma(\nu) = (\nu-1)!$ for positive integer values of ν. Thus, the Gamma function interpolates the factorial function (which is defined only for positive integers). Can we compute it for any other ν? The answer is yes, but only for special values of ν. For example,

$$\Gamma(1/2) = \int_0^\infty x^{-1/2}e^{-x}dx = \sqrt{2}\int_0^\infty e^{-y^2/2}dy$$

by substituting $x = y^2/2$. The last integral was computed above in the context of the normal distribution and equal to $\sqrt{\pi/2}$. Hence, we get $\Gamma(1/2) = \sqrt{\pi}$. From this, using again the relation $\Gamma(\nu+1) = \nu\Gamma(\nu)$, we can compute $\Gamma(3/2) = \frac{1}{2}\sqrt{\pi}$, $\Gamma(5/2) = \frac{3}{4}\sqrt{\pi}$, etc. Yet another useful fact about the Gamma function is its asymptotics as $\nu \to \infty$.

Example 13. Beta distributions: Let $\alpha, \beta > 0$. The Beta distribution with parameters α, β, denoted Beta(α, β), is the distribution with density and distribution given by

$$\text{PDF: } f(t) = \begin{cases} \frac{1}{B(\alpha,\beta)} t^{\alpha-1}(1-t)^{\beta-1} & \text{if } t \in (0,1) \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ \int_0^t f(u)du & \text{if } t \in (0,1) \\ 0 & \text{if } t \geq 1. \end{cases}$$

Here, $B(\alpha, \beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$. Again, for special values of α, β (e.g., positive integers), one can find the value of $B(\alpha, \beta)$, but in general there is no simple expression. However, it can be expressed in terms of the Gamma function!

Proposition 14. *For any $\alpha, \beta > 0$, we have $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.*

Proof. For $\beta = 1$, we see that $B(\alpha, 1) = \int_0^1 t^{\alpha-1} = \frac{1}{\alpha}$ which is also equal to $\frac{\Gamma(\alpha)\Gamma(1)}{\Gamma(\alpha+1)}$ as required. Similarly (or, by the symmetry relation $B(\alpha, \beta) = B(\beta, \alpha)$), we see that $B(1, \beta)$ also has the desired expression.

Now, for any other *positive integer* value of α and real $\beta > 0$ we can integrate by parts and get

$$\begin{aligned} B(\alpha, \beta) &= \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \\ &= -\frac{1}{\beta} t^{\alpha-1}(1-t)^{\beta} \Big|_0^1 + \frac{\alpha-1}{\beta} \int_0^1 t^{\alpha-2}(1-t)^{\beta}dt \\ &= \frac{\alpha-1}{\beta} B(\alpha-1, \beta+1). \end{aligned}$$

Stirling's approximation: $\frac{\Gamma(\nu+1)}{\nu^{\nu+\frac{1}{2}} e^{-\nu} \sqrt{2\pi}} \to 1$ as $\nu \to \infty$.

A small digression: It was Euler's idea to observe that $n! = \int_0^\infty x^n e^{-x} dx$ and that on the right side n could be replaced by any real number greater than -1. But this was his second approach to defining the Gamma function. His first approach was as follows. Fix a positive integer n. Then for any $\ell \geq 1$ (also a positive integer), we may write

$$n! = \frac{(n+\ell)!}{(n+1)(n+2)\cdots(n+\ell)} = \frac{\ell!(\ell+1)\cdots(\ell+n)}{(n+1)\cdots(n+\ell)} = \frac{\ell!\,\ell^n}{(n+1)\cdots(n+\ell)} \cdot \frac{(\ell+1)\cdots(\ell+n)}{\ell^n}$$

The second factor approaches 1 as $\ell \to \infty$. Hence,

$$n! = \lim_{\ell \to \infty} \frac{\ell!\,\ell^n}{(n+1)\cdots(n+\ell)}.$$

Euler then showed (by a rather simple argument that we skip) that the limit on the right exists if we replace n by any complex number other than $\{-1, -2, -3, \ldots\}$ (negative integers are a problem as they make the denominator zero). Thus, he extended the factorial function to all complex numbers except negative integers! It is a fun exercise to check that this agrees with the definition by the integral given earlier. In other words, for $\nu > -1$, we have

$$\lim_{\ell \to \infty} \frac{\ell!\,\ell^\nu}{(\nu+1)\cdots(\nu+\ell)} = \int_0^\infty x^\nu e^{-x} dx.$$

Note that the first term vanishes because $\alpha > 1$ and $\beta > 0$. When α is an integer, we repeat this for α times and get

$$B(\alpha, \beta) = \frac{(\alpha-1)(\alpha-2)\cdots 1}{\beta(\beta+1)\cdots(\beta+\alpha-2)} B(1, \beta+\alpha-1).$$

But, we already checked that $B(1, \beta+\alpha-1) = \frac{\Gamma(1)\Gamma(\alpha+\beta-1)}{\Gamma(\alpha+\beta)}$ from which we get

$$B(\alpha, \beta) = \frac{(\alpha-1)(\alpha-2)\cdots 1}{\beta(\beta+1)\cdots(\beta+\alpha-2)} \frac{\Gamma(1)\Gamma(\alpha+\beta-1)}{\Gamma(\alpha+\beta)} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

by the recursion property of the Gamma function. Thus, we have proved the proposition when α is a positive integer. By symmetry, the same is true when β is a positive integer (and α can take any value). We do not prove the proposition for general $\alpha, \beta > 0$ here. ■

Example 15. The standard Cauchy distribution: A distribution with density and distribution given by

$$\text{PDF: } f(t) = \frac{1}{\pi(1+t^2)} \qquad \text{CDF: } F(t) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1} t.$$

One can also make a parametric family of Cauchy distributions with parameters $\lambda > 0$ and $a \in \mathbb{R}$ denoted Cauchy(a, λ) as follows:

$$\text{PDF: } f(t) = \frac{\lambda}{\pi(\lambda^2 + (t-a)^2)} \qquad \text{CDF: } F(t) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{t-a}{\lambda}\right).$$

Remark 16. Does every CDF come from a pdf? Not necessarily. For example any CDF that is not continuous (for example, CDFs of discrete distributions such as Binomial, Poisson, Geometric etc.). In fact, even continuous CDFs may not have densities (there is a good example manufactured out of the $1/3$-Cantor set, but that would take us out of the topic now). However, suppose F is a *continuous* CDF and suppose F is differentiable except at finitely many points and that the derivative is a continuous function. Then, $f(t) := F'(t)$ defines a pdf which by the fundamental theorm of Calculus satisfies $F(t) = \int_{-\infty}^{t} f(u)du$.

3. Is a CDF necessarily discrete, or continuous?

Let X be a random variable defined on a probability space, and F_X be the distribution function (DF) of X.

Exercise 17. F_X will either be continuous everywhere, or it will have countable number of discontinuities. Moreover, the sum of sizes of jumps at the point of discontinuities of F_X will be either 1, or less than 1.

This property of F_X can be used to classify the random variable X into three broad categories:

- The random variable X is said to be of *discrete* type if there exists a non-empty and countable set S_X such that
$$f_X(x) = \mathbf{P}(\{X = x\}) = F_X(x) - F_X(x-) > 0 \,\forall\, x \in S_X,$$
and $\mathbf{P}\left(S_X\right) = \sum_{x \in S_X} \mathbf{P}(\{X = x\}) = \sum_{x \in S_X} \left[F_X(x) - F_X(x-)\right] = 1$. The function f_X is called the probability mass function (pmf) of the random variable X, and the set S_X is called the support of random variable X (or, of the pmf f_X).
- A random variable X is said to be of *continuous* type if its distribution function F_X is continuous everywhere.
- A random variable X with distribution function F_X is said to be of *absolutely continuous* type if there exists an integrable function $f_X : \mathbb{R} \to \mathbb{R}_+$ such that $f_X(x) \geq 0 \,\forall\, x \in \mathbb{R}$, and
$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \,\forall\, x \in \mathbb{R}.$$
The function f_X is called the probability density function (pdf) of the random variable X, and the set $S_X = \{x \in \mathbb{R} : f_X(x) > 0\}$ is called the support of random variable X (or, of the pdf f_X).

Example 18. Now, consider the following DF:

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \frac{x}{4}, & \text{if } 0 \leq x < 1 \\ \frac{x}{3}, & \text{if } 1 \leq x < 2 \\ \frac{3x}{8}, & \text{if } 2 \leq x < \frac{5}{2} \\ 1, & \text{if } x \geq \frac{5}{2}. \end{cases}$$

Note that the set of discontinuity points of F_X is $\{1, 2, 5/2\}$ with

$$\begin{aligned} p_1 &= \mathbf{P}(\{X = 1\}) = F_X(1) - F_X(1-) = \tfrac{1}{12}, \\ p_2 &= \mathbf{P}(\{X = 2\}) = F_X(2) - F_X(2-) = \tfrac{1}{12} \text{ and} \\ p_3 &= \mathbf{P}(\{X = 5/2\}) = F_X(5/2) - F_X(5/2-) = \tfrac{1}{16}. \end{aligned}$$

Thus, $p_1 + p_2 + p_3 = \frac{11}{48} < 1$.

We can decompose F_X as $F_X(x) = \alpha H_d(x) + (1-\alpha)H_c(x)$ for $x \in \mathbb{R}$, where $\alpha \in [0,1]$. Here, H_d is a distribution function of some random variable X_d of discrete type, while H_c is a distribution function of some random variable X_c of continuous type.

Let us take $\alpha = p_1 + p_2 + p_3 = \frac{11}{48}$. Thus, $\mathbf{P}\left(\{X_d = 1\}\right) = \frac{p_1}{\alpha} = \frac{4}{11}$, $\mathbf{P}\left(\{X_d = 2\}\right) = \frac{p_2}{\alpha} = \frac{4}{11}$ and $\mathbf{P}\left(\{X_d = 5/2\}\right) = \frac{p_3}{\alpha} = \frac{3}{11}$. This gives us

$$H_d(x) = \begin{cases} 0, & \text{if } x < 1 \\ \frac{4}{11}, & \text{if } 1 \le x < 2 \\ \frac{8}{11}, & \text{if } 2 \le x < \frac{5}{2} \\ 1, & \text{if } x \ge \frac{5}{2}, \end{cases}$$

and

$$\begin{aligned} H_c(x) &= \frac{F_X(x) - \alpha H_d(x)}{1-\alpha} \\ &= \begin{cases} 0, & \text{if } x < 0 \\ \frac{12x}{37}, & \text{if } 0 \le x < 1 \\ \frac{4(4x-1)}{37}, & \text{if } 1 \le x < 2 \\ \frac{2(9x-4)}{37}, & \text{if } 2 \le x < \frac{5}{2} \\ 1, & \text{if } x \ge \frac{5}{2}. \end{cases} \end{aligned}$$

Here, the distribution function F_X (equivalently, the random variable X) is neither discrete, nor continuous.

Remark 19. Convex combination of two DFs is also a DF.

4. Change of Variable

Let $h : \mathbb{R} \to \mathbb{R}$ be a function. Given the distribution of X, how will you find the distribution of $h(X)$?

1. CDF technique: The distribution $Z = h(X)$ can be determined by computing the distribution function. Fix $z \in \mathbb{R}$,

$$F_Z(z) = \mathbf{P}(Z \leq z) = \mathbf{P}(h(X) \leq z).$$

Depending on the properties of the function h, we may, or may not be able to derive this probability in a closed form expression.

Example 20. Let X be a random variable with pmf

$$f_X(x) = \begin{cases} \left(\begin{array}{c} n \\ x \end{array}\right) p^x(1-p)^{n-x}, & \text{if } x \in \{0, 1, \ldots, n\} \\ 0, & \text{otherwise ,} \end{cases}$$

where n is a positive integer and $p \in (0,1)$. Find the distribution function of $Y = n - X$.

Note that $S_X = S_Y = \{0, 1, \ldots, n\}$. For $y \in S_Y$, we get

$$\mathbf{P}(\{Y \leq y\}) = \mathbf{P}(\{X \geq n-y\}) = \sum_{x=n-y}^{n} \left(\begin{array}{c} n \\ x \end{array}\right) p^x(1-p)^{n-x} = \sum_{x=0}^{y} \left(\begin{array}{c} n \\ n-x \end{array}\right) (1-p)^x p^{n-x}.$$

Thus, the distribution function of Y is

$$F_Y(y) = \begin{cases} 0, & \text{if } y < 0 \\ p^n, & \text{if } 0 \leq y < 1 \\ \sum_{j=0}^{i} \left(\begin{array}{c} n \\ j \end{array}\right) (1-p)^j p^{n-j}, & \text{if } i \leq y < i+1 \text{ for } i = 1, 2, \ldots, n-1 \\ 1, & \text{if } y \geq n. \end{cases}$$

Example 21. Let X be random variable with pdf

$$f_X(x) = \begin{cases} e^{-x}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

and $T = X^2$. Find the pdf of T.

We have $F_T(t) = \mathbf{P}\left(\{X^2 \leq t\}\right)$ for $t \in \mathbb{R}$. Clearly, $F_T(t) = 0$ for $t < 0$. For $t \geq 0$,

$$\begin{aligned} F_T(t) &= \mathbf{P}(\{-\sqrt{t} \leq X \leq \sqrt{t}\}) \\ &= \int_{-\sqrt{t}}^{\sqrt{t}} f_X(x) \mathrm{d}x \\ &= \int_0^{\sqrt{t}} e^{-x}\, \mathrm{d}x \\ &= 1 - e^{-\sqrt{t}}. \end{aligned}$$

Therefore, the distribution function of T is

$$F_T(t) = \begin{cases} 0, & \text{if } t < 0 \\ 1 - e^{-\sqrt{t}}, & \text{if } t \geq 0. \end{cases}$$

Clearly, F_T is differentiable everywhere except at $t = 0$. Therefore, T is of absolutely continuous type with pdf $f_T(t) = F_T'(t)$ for $t \neq 0$. At $t = 0$, we may assign any arbitrary nonnegative value to $f_T(0)$. Thus, a pdf of T is

$$f_T(t) = \begin{cases} \frac{e^{-\sqrt{t}}}{2\sqrt{t}}, & \text{if } t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Example 22. Let X be a random variable with pdf

$$f_X(x) = \begin{cases} \frac{|x|}{2}, & \text{if } -1 < x < 1 \\ \frac{x}{3}, & \text{if } 1 \leq x < 2 \\ 0, & \text{otherwise} \end{cases}$$

and $T = X^2$. Find the distribution function of T.

We have $F_T(t) = \mathbf{P}\left(\{X^2 \leq t\}\right)$ for $t \in \mathbb{R}$. Since $\mathbf{P}(\{X \in (-1, 2)\}) = 1$, we have $\mathbf{P}(\{T \in (0, 4)\}) = 1$. Therefore, $F_T(t) = \mathbf{P}(\{T \leq t\}) = 0$ for $t < 0$, and $F_T(t) = \mathbf{P}(\{T \leq t\}) = 1$ for $t \geq 4$. For $t \in [0, 4)$, we have

$$\begin{aligned} F_T(t) &= \mathbf{P}(\{-\sqrt{t} \leq X \leq \sqrt{t}\}) \\ &= \int_{-\sqrt{t}}^{\sqrt{t}} f_X(x) \mathrm{d}x \\ &= \begin{cases} \displaystyle\int_{-\sqrt{t}}^{\sqrt{t}} \frac{|x|}{2}\, \mathrm{d}x, & \text{if } 0 \leq t < 1 \\ \displaystyle\int_{-1}^{1} \frac{|x|}{2}\, \mathrm{d}x + \int_1^{\sqrt{t}} \frac{x}{3}\, \mathrm{d}x, & \text{if } 1 \leq t < 4. \end{cases} \end{aligned}$$

Therefore, the distribution function of T is

$$F_T(t) = \begin{cases} 0, & \text{if } t < 0 \\ \frac{t}{2}, & \text{if } 0 \leq t < 1 \\ \frac{t+2}{6}, & \text{if } 1 \leq t < 4 \\ 1, & \text{if } t \geq 4. \end{cases}$$

Clearly, F_T is differentiable everywhere except at points 0, 1 and 4. It follows that the random variable T is of absolutely continuous type with pdf

$$f_T(t) = \begin{cases} \frac{1}{2}, & \text{if } 0 < t < 1 \\ \frac{1}{6}, & \text{if } 1 < t < 4 \\ 0, & \text{otherwise.} \end{cases}$$

2.a. Change of variable (for discrete probability distributions): Let X be a random variable of discrete type with support S_X and pmf f_X. Define $Z = h(X)$. Then, Z is a random variable of discrete type with support $S_Z = \{h(x) : x \in S_X\}$ with pmf

$$f_Z(z) = \begin{cases} \sum_{x \in A_z} f_X(x), & \text{if } z \in S_Z \\ 0, & \text{otherwise,} \end{cases}$$

where $A_z = \{x \in S_X : h(x) = z\}$.

Corollary: Suppose that $h : \mathbb{R} \to \mathbb{R}$ is one-one with inverse function $h^{-1} : D \to \mathbb{R}$, where $D = \{h(x) : x \in \mathbb{R}\}$. Then, Z is a discrete type random variable with support $S_Z = \{h(x) : x \in S_X\}$ and pmf

$$f_Z(z) = \begin{cases} f_X(h^{-1}(z)), & \text{if } z \in S_Z \\ 0, & \text{otherwise.} \end{cases}$$

Example 23. Let X be a random variable with pmf

$$f_X(x) = \begin{cases} \frac{1}{7}, & \text{if } x \in \{-2, -1, 0, 1\} \\ \frac{3}{14}, & \text{if } x \in \{2, 3\} \\ 0, & \text{otherwise.} \end{cases}$$

Find the pmf and distribution function of $Z = X^2$.

Clearly, $S_X = \{-2, -1, 0, 1, 2, 3\}$ and $S_Z = \{0, 1, 4, 9\}$. Moreover,

$$\begin{aligned} &\mathbf{P}(\{Z = 0\}) = \mathbf{P}\left(\{X^2 = 0\}\right) = \mathbf{P}(\{X = 0\}) = \tfrac{1}{7}, \\ &\mathbf{P}(\{Z = 1\}) = \mathbf{P}\left(\{X^2 = 1\}\right) = \mathbf{P}(X \in \{-1, 1\}) = \tfrac{1}{7} + \tfrac{1}{7} = \tfrac{2}{7}, \\ &\mathbf{P}(\{Z = 4\}) = \mathbf{P}\left(\{X^2 = 4\}\right) = \mathbf{P}(X \in \{-2, 2\}) = \tfrac{1}{7} + \tfrac{3}{14} = \tfrac{5}{14} \text{ and} \\ &\mathbf{P}(\{Z = 9\}) = \mathbf{P}\left(\{X^2 = 9\}\right) = \mathbf{P}(X \in \{-3, 3\}) = 0 + \tfrac{3}{14} = \tfrac{3}{14}. \end{aligned}$$

Therefore, the pmf of Z is

$$f_Z(z) = \begin{cases} \frac{1}{7}, & \text{if } z = 0 \\ \frac{2}{7}, & \text{if } z = 1 \\ \frac{5}{14}, & \text{if } z = 4 \\ \frac{3}{14}, & \text{if } z = 9 \\ 0, & \text{otherwise,} \end{cases}$$

and the distribution function of Z is

$$F_Z(z) = \begin{cases} 0, & \text{if } z < 0 \\ \frac{1}{7}, & \text{if } 0 \leq z < 1 \\ \frac{3}{7}, & \text{if } 1 \leq z < 4 \\ \frac{11}{14}, & \text{if } 4 \leq z < 9 \\ 1, & \text{if } z \geq 9. \end{cases}$$

Example 24. Let X be a random variable with pmf

$$f_X(x) = \begin{cases} \frac{|x|}{2550} & \text{if } x \in \{\pm 1, \pm 2, \dots, \pm 50\} \\ 0, & \text{otherwise.} \end{cases}$$

Find the pmf and distribution function of $Z = |X|$.

We have $S_X = \{\pm 1, \pm 2, \dots, \pm 50\}$ and $S_Z = \{1, 2, \dots, 50\}$. Moreover, for $z \in S_Z$

$$\mathbf{P}(\{Z = z\}) = \mathbf{P}(\{|X| = z\}) = \mathbf{P}(\{X \in \{-z, z\}\}) = \frac{|-z|}{2550} + \frac{|z|}{2550} = \frac{z}{1275}.$$

Therefore, the pmf of Z is

$$f_Z(z) = \begin{cases} \frac{z}{1275}, & \text{if } z \in \{1, 2, \dots, 50\} \\ 0, & \text{otherwise,} \end{cases}$$

and the distribution function of Z is

$$F_Z(z) = \begin{cases} 0, & \text{if } z < 1 \\ \frac{1}{1275}, & \text{if } 1 \leq z < 2 \\ \frac{i(i+1)}{2550}, & \text{if } i \leq z < i+1 \text{ for } i = 2, 3, \dots, 49 \\ 1, & \text{if } z \geq 50. \end{cases}$$

Example 25. Let X be a random variable with pmf

$$f_X(x) = \begin{cases} \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}, & \text{if } x \in \{0, 1, \dots, n\} \\ 0, & \text{otherwise,} \end{cases}$$

where n is a positive integer and $p \in (0, 1)$. Find the pmf and distribution function of $Y = n - X$.

Note that $S_X = S_Y = \{0, 1, \ldots, n\}$. For $y \in S_Y$, we get

$$\mathbf{P}(\{Y = y\}) = \mathbf{P}(\{X = n - y\}) = \left(\begin{array}{c} n \\ n-y \end{array} \right) p^{n-y}(1-p)^y = \left(\begin{array}{c} n \\ y \end{array} \right) (1-p)^y p^{n-y}.$$

Thus, the pmf of Y is

$$f_Y(y) = \begin{cases} \left(\begin{array}{c} n \\ y \end{array} \right) (1-p)^y p^{n-y}, & \text{if } y \in \{0, 1, \ldots, n\} \\ 0, & \text{otherwise,} \end{cases}$$

and the distribution function of Y is

$$F_Y(y) = \begin{cases} 0, & \text{if } y < 0 \\ p^n, & \text{if } 0 \le y < 1 \\ \sum_{j=0}^{i} \left(\begin{array}{c} n \\ j \end{array} \right) (1-p)^j p^{n-j}, & \text{if } i \le y < i+1 \text{ for } i = 1, 2, \ldots, n-1 \\ 1, & \text{if } y \ge n. \end{cases}$$

2.b. Change of variable (for continuous probability distributions): Let X be a random variable of absolutely continuous type with pdf f_X and support S_X. Let $S_1, S_2, \ldots, S_k$ be open intervals in $\mathbb{R}$ such that $S_i \cap S_j = \emptyset$ if $i \neq j$ and $\cup_{i=1}^{k} S_i = S_X$.

Let $h : \mathbb{R} \to \mathbb{R}$ be a function such that on each S_i, the function $h : S_j \to \mathbb{R}$ is *strictly monotone* and *continuously differentiable* with inverse function (say, h_j^{-1}) for $j = 1, \ldots, k$. Let $h(S_j) = \{h(x) : x \in S_j\}$ so that $h(S_j)$ is an open interval in $\mathbb{R}$ for $j = 1, \ldots, k$.

Then, the random variable $T = h(X)$ is continuous with pdf

$$f_T(t) = \sum_{j=1}^{k} f_X\left(h_j^{-1}(t)\right) \left| \frac{\mathrm{d}}{\mathrm{d}t} h_j^{-1}(t) \right| I_{h(S_j)}(t).$$

Corollary: Let X be a continuous random variable with pdf f_X and support S_X. Suppose that S_X is a finite union of disjoint open intervals in $\mathbb{R}$, and let $h : \mathbb{R} \to \mathbb{R}$ be *differentiable and strictly monotone* on S_X (i.e., either $h'(x) < 0 \; \forall \; x \in S_X$ or $h'(x) > 0 \; \forall \; x \in S_X$). Let $S_T = \{h(x) : x \in S_X\}$. Then, $T = h(X)$ is a continuous random variable with pdf

$$f_T(t) = \begin{cases} f_X(h^{-1}(t)) \left| \frac{\mathrm{d}}{\mathrm{d}t} h^{-1}(t) \right|, & \text{if } t \in S_T \\ 0, & \text{otherwise.} \end{cases}$$

Example 26. Let X be random variable with pdf

$$f_X(x) = \begin{cases} e^{-x}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

and $T = X^2$. Find the distribution function of T, and hence find its pdf.

We have $S_X = S_T = (0, \infty)$. Also, $h(x) = x^2$ for $x \in S_X$ is strictly increasing on S_X with inverse function $h^{-1}(x) = \sqrt{x}$ for $x \in S_T$. It follows that $T = X^2$ has pdf

$$f_T(t) = \begin{cases} f_X(\sqrt{t}) \left| \frac{d}{dt}(\sqrt{t}) \right|, & \text{if } t > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{e^{-\sqrt{t}}}{2\sqrt{t}}, & \text{if } t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Example 27. Let X be a random variable with pdf

$$f_X(x) = \begin{cases} \frac{|x|}{2}, & \text{if } -1 < x < 1 \\ \frac{x}{3}, & \text{if } 1 \leq x < 2 \\ 0, & \text{otherwise} \end{cases}$$

and $T = X^2$. Find the distribution function of T, and hence find its pdf.

We have $S_X = (-1, 0) \cup (0, 2) = S_1 \cup S_2$, say. Now, $h(x) = x^2$ is strictly decreasing in $S_1 = (-1, 0)$ with inverse function $h_1^{-1}(t) = -\sqrt{t}$; and $h(x) = x^2$ is strictly increasing in $S_2 = (0, 2)$ with inverse function $h_2^{-1}(t) = \sqrt{t}$. Note that $h(S_1) = (0, 1)$ and $h(S_2) = (0, 4)$. It now follows that $T = X^2$ has pdf

$$f_T(t) = f_X(-\sqrt{t}) \left| \frac{\mathrm{d}}{\mathrm{d}t}(-\sqrt{t}) \right| I_{(0,1)}(t) + f_X(\sqrt{t}) \left| \frac{\mathrm{d}}{\mathrm{d}t}(\sqrt{t}) \right| I_{(0,4)}(t)$$

$$= \begin{cases} \frac{1}{2}, & \text{if } 0 < t < 1 \\ \frac{1}{6}, & \text{if } 1 < t < 4 \\ 0, & \text{otherwise.} \end{cases}$$

Example 28. Let X be a random variable of absolutely continuous type with pdf

$$f_X(x) = \begin{cases} e^{-x}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

and $T = [X]$, where $[x]$ denotes the largest integer not exceeding x for $x \in \mathbb{R}$. Find its pmf.

Note that $S_X = (0, \infty)$. Since $\mathbf{P}\left(\{X \in S_X\}\right) = 1$, we have $\mathbf{P}(T \in \{0, 1, 2, \ldots\}) = 1$. Also, for $i \in \{0, 1, 2, \ldots\}$

$$\begin{aligned}\mathbf{P}(\{T = i\}) &= \mathbf{P}(\{i \leq X < i+1\}) \\ &= \int_i^{i+1} f_X(x)\mathrm{d}x \\ &= \int_i^{i+1} e^{-x}\,\mathrm{d}x \\ &= \left(1 - e^{-1}\right) e^{-i}.\end{aligned}$$

Consequently, the random variable T is of discrete type with support $S_T = \{0, 1, 2, \ldots\}$ with pmf

$$f_T(t) = \mathbf{P}(\{T = t\}) = \begin{cases} \left(1 - e^{-1}\right) e^{-t}, & \text{if } t \in \{0, 1, 2, \ldots\} \\ 0, & \text{otherwise.} \end{cases}$$

Remark 29. This example illustrates that in general, a function of an absolutely continuous type random variable may not be of absolutely continuous type.

1. Mean and Variance

Let X be a random variable with distribution F_X. We shall assume that it has pmf (or, pdf) denoted by f_X.

Definition 1. The *expected value* (also called *mean*) of X is defined as the quantity $\mathbf{E}[X] = \sum_t tf(t)$ if f is a pmf, and $\mathbf{E}[X] = \int_{-\infty}^{+\infty} tf(t)dt$ if f is a pdf (provided the sum, or the integral converges absolutely).

In other words,

$$\mathbf{E}[X] = \begin{cases} \sum_x xf(x), & \text{if } \sum_x |x|f(x) < \infty \text{ for discrete } X, \\ \int_{-\infty}^{+\infty} xf(x), & \text{if } \int_{-\infty}^{+\infty} |x|f(x) < \infty \text{ for continuous } X. \end{cases}$$

Note that it is possible to define expected value for distributions without pmf or pdf, but we shall not do it here.

Exercise 2. Find the expectation of random variables with the following pdf/pmf:

(a) $f(x) = \frac{1}{\pi(1+x^2)}$ when $x \in \mathbb{R}$,

(b) $f(x) = \frac{1}{|x|(1+|x|)}$ when $x \in S = \{(-1)^n n | n \in \mathbb{N}\}$.

Properties of expectation: Let X, Y be random variables both having pmf f, g (or, pdf f, g).

(1) Then, $\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y]$ for any $a, b \in \mathbb{R}$. In particular, for a constant random variable (i.e., $X = a$ with probability 1 for some a, $\mathbf{E}[X] = a$). This is called *linearity* of expectation.

(2) If $X \geq Y$ (meaning, $X(\omega) \geq Y(\omega)$ for all ω), then $\mathbf{E}[X] \geq \mathbf{E}[Y]$.

(3) If $\varphi : \mathbb{R} \to \mathbb{R}$, then

$$\mathbf{E}[\varphi(X)] = \begin{cases} \sum\limits_t \varphi(t)f(t) & \text{if } f \text{ is a pmf} \\ \int_{-\infty}^{+\infty} \varphi(t)f(t)dt & \text{if } f \text{ is a pdf,} \end{cases}$$

provided they exist (i.e., the sum, or the integral converges absolutely).

For random variables on a discrete probability space (they have a pmf), we can easily prove all these properties. For random variables with pdf, a proper proof require a bit of work.

Note: Expectation is a very important quantity. Using it, we can define several other quantities of interest.

Discussion: For simplicity, let us take random variables to have densities in this discussion. You may adapt the remarks to the case of pmf easily. The density has all the information we need about a random variable. However, it is a function, which means that we have to know $f(t)$ for every t. In real life, we often have random variables whose pdf is unknown, or impossible to determine. It would be better to summarize the main features of the distribution (i.e., the density) in a few numbers. That is what the quantities defined below try to do.

Mean: Mean is another term for expected value.

Moments: The quantity $\mu'_k = \mathbf{E}[X^k]$ (if it exists) is called the k^{th} *moment* of X for $k \in \{1, 2, \ldots\}$.

Central Moments: The quantity $\mu_k = \mathbf{E}[(X - \mu)^k]$ (if it exists) is called the k^{th} *central moment* of X for $k \in \{1, 2, \ldots\}$. Here, and henceforth $\mu = \mu'_1$ denotes the mean.

Variance: Let $\mu = \mathbf{E}[X]$ and define $\sigma^2 := \mathbf{E}\left[(X - \mu)^2\right]$. This is called the *variance* of X, also denoted by $\text{Var}(X)$. It can be written in other forms. For example,

$$\begin{aligned}\sigma^2 &= \mathbf{E}[X^2 + \mu^2 - 2\mu X] \qquad \text{(by expanding the square)}\\ &= \mathbf{E}[X^2] + \mu^2 - 2\mu\mathbf{E}[X] \qquad \text{(by property (1) above)}\\ &= \mathbf{E}[X^2] - \mu^2.\end{aligned}$$

That is $\text{Var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$.

Standard deviation: The standard deviation of X is defined as $\text{s.d.}(X) := \sqrt{\text{Var}(X)}$.

Computing expectations from the pmf: Let X be a random variable on (Ω, p) with pmf f. Then, we claim that

$$\mathbf{E}[X] = \sum_{t \in \mathbb{R}} t f(t).$$

Indeed, let $\text{Range}(X) = \{x_1, x_2, \ldots\}$. Let $A_k = \{\omega : X(\omega) = x_k\}$. By the definition of pmf, we have $\mathbf{P}(A_k) = f(x_k)$. Further, A_k are pairwise disjoint and exhaustive. Hence

$$\mathbf{E}[X] = \sum_{\omega \in \Omega} X(\omega) p_\omega = \sum_k \sum_{\omega \in A_k} X(\omega) p_\omega = \sum_k x_k \mathbf{P}(A_k) = \sum_k x_k f(x_k).$$

Similarly, $\mathbf{E}[X^2] = \sum_k x_k^2 f(x_k)$.

Remark 3. More generally, if $h : \mathbb{R} \to \mathbb{R}$ is any function, then the random variable $h(X)$ has expectation $\mathbf{E}[h(X)] = \sum_k h(x_k) f(x_k)$. Although this sounds trivial, there is a very useful point here. To calculate $\mathbf{E}[X^2]$ we do not have to compute the pmf of X^2 first, which can be done but would be more complicated. Instead, in the above formulas, $\mathbf{E}[h(X)]$ has been computed directly in terms of the pmf of X.

Exercise 4. Find $\mathbf{E}[X]$, $\mathbf{E}[X^2]$ and $\text{Var}(X)$ in each case.

(1) $X \sim \text{Bin}(n,p)$.

(2) $X \sim \text{Geo}(p)$.

(3) $X \sim \text{Pois}(\lambda)$.

(4) $X \sim \text{Hypergeo}(b,w,m)$.

Example 5. Let $X \sim N(\mu,\sigma^2)$. Recall that its density is $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. We can compute

$$\mathbf{E}[X] = \frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} xe^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = \mu.$$

On the other hand

$$\begin{aligned}\text{Var}(X) &= \frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}(x-\mu)^2e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \;=\; \sigma^2\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}u^2e^{-\frac{u^2}{2}}du \qquad (\text{substitute } x=\mu+\sigma u)\\ &= \sigma^2\frac{2}{\sqrt{2\pi}}\int\limits_{0}^{+\infty}u^2e^{-\frac{u^2}{2}}du \;=\; \sigma^2\frac{2\sqrt{2}}{\sqrt{2\pi}}\int\limits_{0}^{+\infty}\sqrt{t}e^{-t}dt \qquad (\text{substitute } t=u^2/2)\\ &= \sigma^2\frac{2\sqrt{2}}{\sqrt{2\pi}}\Gamma(3/2) \;=\; \sigma^2.\end{aligned}$$

To get the last line, observe that $\Gamma(3/2) = \frac{1}{2}\Gamma(1/2)$ and $\Gamma(1/2) = \sqrt{\pi}$. Thus, we now have a meaning for the parameters μ and σ^2 - they are the mean and variance of the $N(\mu,\sigma^2)$ distribution. Again, note that the mean is the same for all $N(0,\sigma^2)$ distributions but the variances are different, capturing the spread of the distribution.

Exercise 6. Let $X \sim N(0,1)$. Show that $\mathbf{E}[X^n] = 0$ if n is odd. If n is even, then $\mathbf{E}[X^n] = (n-1)(n-3)\cdots(3)(1)$ (product of all odd numbers upto and including $n-1$). What happens if $X \sim N(0,\sigma^2)$?

Exercise 7. If X is a non-negative integer-valued random variable with finite expectation, then

$$\mathbf{E}(X) = \sum_{k=1}^{\infty} \mathbf{P}(X \geq k)$$

Exercise 8. Find $\mathbf{E}[X]$, $\mathbf{E}[X^2]$ and $\mathrm{Var}(X)$ in each case.

(1) $X \sim \mathrm{Exp}(\lambda)$.

(2) $X \sim \mathrm{Gamma}(\nu, \lambda)$.

(3) $X \sim \mathrm{Unif}([0, 1])$.

(4) $X \sim \mathrm{Beta}(p, q)$.

Exercise 9. Show that $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$ (hence, $\mathrm{s.d.}(cX) = |c|\mathrm{s.d.}(X)$) for $c \in \mathbb{R}$.

2. Descriptive Measures of Probability Distributions

Mean: Recall that the mean of a random variable (probability distribution) X is given by $\mu = \mathbf{E}(X)$.

Median: A real number m satisfying

$$F_X(m-) \leq \frac{1}{2} \leq F_X(m), \text{ i.e., } \mathbf{P}(\{X < m\}) \leq \frac{1}{2} \leq \mathbf{P}(\{X \leq m\})$$

is called the median (of the probability distribution) of X.

Note: Let us assume that the CDF F_X of X is *strictly increasing and continuous*. Then $F_X^{-1}(t)$ is well defined for every $t \in (0,1)$. For each $t \in (0,1)$, the number $Q_t = F_X^{-1}(t)$ is called the t-quantile. For example, the $1/2$-quantile, also called *median* is the number x such that $F_X(x) = \frac{1}{2}$ (unique when the CDF is strictly increasing and continuous). Similarly, one defines $1/4$-quantile and $3/4$-quantile and these are sometimes called quartiles.[1]

Mode: The mode m_0 of a probability distribution is the value that occurs with highest pmf/pdf, and is defined by $f_X(m_0) = \sup\{f_X(x) : x \in S_X\}$.

Measures of central tendency: Mean and median try to summarize the distribution of X by a single number. Of course one number cannot capture the whole distribution, so there are many densities and mass functions that have the same mean or median. Which is better - mean or median? This question has no unambiguous answer. Mean has excellent mathematical properties (mainly linearity) which the median lacks (med$(X+Y)$ bears no general relationship to med$(X)+$ med(Y)). In contrast, mean is sensitive to outliers, while the median is far less so. For example, if the average income in a village of 50 people is Rs.1000 per month, the immigration of multi-millionaire to the village will change the mean drastically, but the median remains about the same. This is good, if by giving one number we are hoping to express the state of a typical individual in the population.

[1] Another familiar quantity is the percentile, frequently used in reporting performance in competitive exams. For each x, the x-percentile is nothing but $F(x)$. For exam scores, it tells the proportion of exam-takers who scored less than or equal to x.

Coefficient of variation: The coefficient of variation of X is defined as $\text{c.v.}(X) = \frac{\text{s.d.}(X)}{|\mathbf{E}[X]|}$.

Mean absolute deviation (m.a.d.): The mean absolute deviation of X is defined as the $\mathbf{E}[|X - \text{med}(X)|]$.

Quartile deviation: Let q_1 and q_3 be real numbers such that

$$F_X(q_1-) \leq \frac{1}{4} \leq F_X(q_1) \quad \text{and} \quad F_X(q_3-) \leq \frac{3}{4} \leq F_X(q_3)$$

i.e., $\mathbf{P}(\{X < q_1\}) \leq \frac{1}{4} \leq \mathbf{P}(\{X \leq q_1\}) \quad \text{and} \quad \mathbf{P}(\{X < q_3\}) \leq \frac{3}{4} \leq \mathbf{P}(\{X \leq q_3\})$. The quantities q_1 and q_3 are called, respectively, the lower and upper quartiles of the probability distribution of random variable X.

The quartile deviation (or, inter-quartile range) is defined as $q_3 - q_1$.

Measures of dispersion: Suppose the average height of people in a city is 160 cm. This could be because everyone is 160 cm exactly, or because half the people are 100 cm. While the other half are 220 cm., or alternately the heights could be uniformly spread over 150-170 cm., etc. How widely the distribution is spread is measured by standard deviation and mean absolute deviation. Since we want deviation from mean, $\mathbf{E}[X - \mathbf{E}[X]]$ looks natural, but this is zero because of cancellation of positive and negative deviations. To prevent cancellation, we may put absolute values (getting to the m.a.d, but that is usually taken around the median) or we may square the deviations before taking expectation (giving the variance, and then the standard deviation). Variance and standard deviation have much better mathematical properties, and hence are usually preferred.

The standard deviation has the same units as the quantity. For example, if mean height is 160cm measured in centimeters with a standard deviation of 10cm, and the mean weight is 55kg with a standard deviation of 5kg, then we cannot say which of the two is less variable. To make such a comparison we need a dimension free quantity (a pure number). Coefficient of variation is such a quantity, as it measure the standard deviation per mean. For the height and weight data just described, the coefficients of variation are $1/16$ and $1/11$, respectively. Hence, we may say that height is less variable than weight in this example.

Standardization: Let $\mu = \mathbf{E}[X]$ and $\sigma = \text{s.d.}(X)$, i.e., the mean and the standard deviation of X, respectively. Define $Z = \frac{(X-\mu)}{\sigma}$ to be the standardized variable (independent of units).

Skewness: A measure of skewness of the probability distribution of X is defined by

$$\beta_1 = \mathbf{E}\left(Z^3\right) = \frac{\mathbf{E}\left((X-\mu)^3\right)}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}.$$

Kurtosis: A measure of kurtosis of the probability distribution of X is defined by

$$\gamma_1 = \mathbf{E}\left(Z^4\right) = \frac{\mathbf{E}\left((X-\mu)^4\right)}{\sigma^4} = \frac{\mu_4}{\mu_2^2}.$$

The quantity γ_1 is simply called the kurtosis of the probability distribution of X. It is easy to show that for any values of $\mu \in \mathbb{R}$ and $\sigma > 0$, the kurtosis of $N\left(\mu, \sigma^2\right)$ distribution is $\gamma_1 = 3$. The quantity $\gamma_2 = \gamma_1 - 3$ is called the *excess kurtosis* of the distribution of X.

Exercise 10. Show that for any values of $\mu \in \mathbb{R}$ and $\sigma > 0$, the kurtosis of $N\left(\mu, \sigma^2\right)$ distribution is $\gamma_1 = 3$. This implies that $\gamma_2 = 0$.

3. Moment Generating Function, and its properties

Moment generating function: We are familiar with the Laplace transform of a given real-valued function defined on $\mathbb{R}$. We also know that under certain conditions, the Laplace transform of a function determines the function almost uniquely. In probability theory, the Laplace transform of a pdf/pmf of a random variable X plays an important role and is referred to as moment generating function (of probability distribution) of the random variable X.

Define $M_X : A \to \mathbb{R}$ by

$$M_X(t) = \mathbf{E}[e^{tX}], \quad t \in A.$$

We call M_X the moment generating function (mgf) of the random variable X. We say that the mgf of a random variable X exists if there exists a positive real number a such that $(-a, a) \subseteq A$ (i.e., if $M_X(t) = \mathbf{E}[e^{tX}]$ is finite in an interval containing 0).

Note that $M_X(0) = 1$, and therefore, $A = \{t \in \mathbb{R} : \mathbf{E}[e^{tX}] \text{ is finite}\} \neq \emptyset$. Moreover, we have $M_X(t) > 0 \; \forall \; t \in A$. The name moment generating function to the transform M_X is derived from the fact that M_X can be used to generate moments of random variable X. Let X be a random variable with mgf M_X, which is finite on an interval $(-a, a)$, for some $a > 0$ (i.e., mgf of X exists). Then, we have the following:

(i) $\mu_r' = \mathbf{E}[X^r]$ is finite for each $r \in \{1, 2, \ldots\}$,

(ii) $\mu_r' = \mathbf{E}[X^r] = M_X^{(r)}(0)$, where $M_X^{(r)}(0) = \left[\frac{d^r}{dt^r} M_X(t)\right]_{t=0}$ the r-th derivative of $M_X(t)$ at the point 0 for each $r \in \{1, 2, \ldots\}$, and

(iii) $M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu_r'$ with $t \in (-a, a)$.

Proposition 11. *Under the notation and assumptions of the theorem define $\psi_X : (-a, a) \to \mathbb{R}$ by $\psi_X(t) = \ln M_X(t), t \in (-a, a)$. Then*

$$\mu_1' = \mathbf{E}[X] = \psi_X^{(1)}(0) \quad \textit{and} \quad \mu_2 = \textit{Var}(X) = \psi_X^{(2)}(0),$$

where $\psi_X^{(r)}$ denotes the r-th $(r \in \{1, 2\})$ derivative of ψ_X.

Proof. We have, for $t \in (-a, a)$

$$\psi_X^{(1)}(t) = \frac{M_X^{(1)}(t)}{M_X(t)} \quad \text{and} \quad \psi_X^{(2)}(t) = \frac{M_X(t)M_X^{(2)}(t) - \left(M_X^{(1)}(t)\right)^2}{(M_X(t))^2}.$$

■

Example 12. Let X be a random variable with pmf

$$f_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!}, & \text{if } x \in \{0,1,2,\ldots\} \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$.
(i) Find the mgf $M_X(t)$ for $t \in A = \left\{s \in \mathbb{R} : \mathbf{E}[e^{sX}] < \infty\right\}$ of X. Show that X possesses moments of all orders. Find the mean and variance of X;
(ii) Find $\psi_X(t) = \ln\left(M_X(t)\right)$ for $t \in A$. Hence, find the mean and variance of X;
(iii) What are the first four terms in the power series expansion of M_X around the point 0?

(i) We have

$$M_X(t) = \mathbf{E}\left[e^{tX}\right] = \sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\left(\lambda e^t\right)^x}{x!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda\left(e^t-1\right)}\ \forall\, t \in \mathbb{R},$$

since $A = \left\{s \in \mathbb{R} : \mathbf{E}\left(e^{sX}\right) < \infty\right\} = \mathbb{R}$. For every $r \in \{1,2,\ldots\}$, $\mu_r' = \mathbf{E}\left(X^r\right)$ is finite. Clearly,

$$M_X^{(1)}(t) = \lambda e^t e^{\lambda\left(e^t-1\right)} \quad \text{and} \quad M_X^{(2)}(t) = \lambda e^t e^{\lambda\left(e^t-1\right)}\left(1+\lambda e^t\right)\ \forall\, t \in \mathbb{R}.$$

Therefore,

$$\mathbf{E}(X) = M_X^{(1)}(0) = \lambda,$$
$$\mathbf{E}\left(X^2\right) = M_X^{(2)}(0) = \lambda(1+\lambda) \text{ and}$$
$$\text{Var}(X) = \mathbf{E}\left(X^2\right) - (\mathbf{E}(X))^2 = \lambda.$$

(ii) We have, for $t \in \mathbb{R}$

$$\psi_X(t) = \ln\left(M_X(t)\right) = \lambda\left(e^t - 1\right) \Rightarrow \quad \psi_X^{(1)}(t) = \psi_X^{(2)}(t) = \lambda e^t.$$

Therefore,

$$\mathbf{E}(X) = \psi_X^{(1)}(0) = \lambda \quad \text{and} \quad \text{Var}(X) = \psi_X^{(2)}(0) = \lambda.$$

(iii) We have

$$\begin{aligned} &M_X^{(3)}(t) = \lambda e^t e^{\lambda\left(e^t-1\right)}\left(\lambda^2 e^{2t} + 3\lambda e^t + 1\right)\ \forall\, t \in \mathbb{R} \\ \Rightarrow \quad &\mu_3' = \mathbf{E}\left(X^3\right) = M_X^{(3)}(0) = \lambda\left(\lambda^2 + 3\lambda + 1\right). \end{aligned}$$

Since $A = \left\{s \in \mathbb{R} : \mathbf{E}\left(e^{sX}\right) < \infty\right\} = \mathbb{R}$, we have

$$\begin{aligned} M_X(t) &= 1 + \mu_1' t + \mu_2'\frac{t^2}{2!} + \mu_3'\frac{t^3}{3!} + \cdots \\ &= 1 + \lambda t + \lambda(\lambda+1)\frac{t^2}{2!} + \lambda\left(\lambda^2 + 3\lambda + 1\right)\frac{t^3}{3!} + \cdots, t \in \mathbb{R}. \end{aligned}$$

Example 13. Let X be a random variable with pdf

$$f_X(x) = \begin{cases} e^{-x}, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(i) Find the mgf $M_X(t)$ for $t \in A = \left\{s \in \mathbb{R} : \mathbf{E}\left(e^{sX}\right) < \infty\right\}$ of X. Show that X possesses moments of all orders. Find the mean and variance of X.
(ii) Find $\psi_X(t) = \ln\left(M_X(t)\right)$ for $t \in A$. Hence, find the mean and variance of X.
(iii) Expand M_X as a power series around the point 0 and hence find $\mathbf{E}\left(X^r\right)$ with $r \in \{1, 2, \ldots\}$.

(i) We have

$$M_X(t) = E\left(e^{tX}\right) = \int_0^\infty e^{tx} e^{-x}\, \mathrm{d}x = \int_0^\infty e^{-(1-t)x}\, \mathrm{d}x < \infty, \text{ if } t < 1.$$

Clearly, $A = \left\{s \in \mathbb{R} : \mathbf{E}\left(e^{sX}\right) < \infty\right\} = (-\infty, 1) \supset (-1, 1)$ and $M_X(t) = (1-t)^{-1}$ for $t < 1$. For every $r \in \{1, 2, \ldots\}$, μ_r' is finite. Clearly,

$$M_X^{(1)}(t) = (1-t)^{-2} \quad \text{and} \quad M_X^{(2)}(t) = 2(1-t)^{-3} \text{ for } t < 1.$$

So, we get

$$\mathbf{E}(X) = M_X^{(1)}(0) = 1,$$
$$\mathbf{E}\left(X^2\right) = M_X^{(2)}(0) = 2 \text{ and}$$
$$\mathrm{Var}(X) = \mathbf{E}\left(X^2\right) - (\mathbf{E}(X))^2 = 1.$$

(ii) For $t < 1$, we have

$$\psi_X(t) = \ln\left(M_X(t)\right) = -\ln(1-t) \Rightarrow \quad \psi_X^{(1)}(t) = \frac{1}{1-t} \text{ and } \psi_X^{(2)}(t) = \frac{1}{(1-t)^2}.$$

So, we get

$$\mathbf{E}(X) = \psi_X^{(1)}(0) = 1 \quad \text{and} \quad \mathrm{Var}(X) = \psi_X^{(2)}(0) = 1.$$

(iii) We have

$$M_X(t) = (1-t)^{-1} = \sum_{r=0}^{\infty} t^r, \quad \text{for } t \in (-1, 1),$$

since $A = \left\{s \in \mathbb{R} : \mathbf{E}\left(e^{sX}\right) < \infty\right\} = (-\infty, 1) \supset (-1, 1)$. So, we conclude that $\mu_r' =$ coefficient of $\frac{t^r}{r!}$ in the power series expansion of M_X around 0.

Exercise 14. Let X be a random variable with pdf

$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, -\infty < x < \infty$$

Show that the mgf of X does not exist.

Identically Distributed Random Variables: Two random variables X and Y are said to have the same distribution (written as $X \overset{D}{=} Y$) if they have the same distribution function, i.e., if

$$F_X(x) = F_Y(x) \; \forall \; x \in \mathbb{R}.$$

(i) Let X and Y be random variables of discrete type with pmf f_X and f_Y, respectively. Then $X \overset{D}{=} Y$ if and only if $f_X(x) = f_Y(x) \; \forall \; x \in \mathbb{R}$.

(ii) Let X and Y be random variables of continuous type with pdf f_X and f_Y, respectively. Then $X \overset{D}{=} Y$ if and only if $f_X(x) = f_Y(x) \; \forall \; x \in \mathbb{R}$.

(iii) Let X and Y be random variables having mgfs M_X and M_Y, respectively. Suppose that there exists a positive real number b such that $M_X(t) = M_Y(t) \; \forall \; t \in (-b, b)$. Then, $X \overset{D}{=} Y$.

Remark 15. The idea of two random variables being identical in distribution is different from two random variables being identical (equal) pointwise. Which notion is stronger?

Symmetric Distribution: A random variable X is said to have a symmetric distribution about a point $\mu \in \mathbb{R}$ if $X - \mu \overset{D}{=} \mu - X$.

Exercise 16. Check that the $N(\mu, \sigma^2)$ distribution is symmetric about the point μ. Find the point of symmetry of Bin$(n, 1/2)$ distrbution, when n is a positive, even integer.

4. Markov's and Chebyshev's inequalities

Let X be a non-negative integer valued random variable with pmf $f(k)$ for $k = 0, 1, 2, \ldots$. Fix any number m, say $m = 10$. Then

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} 10f(k) = 10\mathbf{P}\{X \geq 10\}.$$

More generally, $m\mathbf{P}\{X \geq m\} \leq \mathbf{E}[X]$.

Markov's inequality: Let X be a non-negative random variable with finite expectation. Then, for any $t > 0$, we have $\mathbf{P}\{X \geq t\} \leq \frac{1}{t}\mathbf{E}[X]$.

Proof. Fix $t > 0$ and let $Y = X\mathbf{1}_{X<t}$ and $Z = X\mathbf{1}_{X \geq t}$ so that $X = Y + Z$. Both Y and Z are non-negative random variable and hence $\mathbf{E}[X] = \mathbf{E}[Y] + \mathbf{E}[Z] \geq \mathbf{E}[Z]$. On the other hand, $Z \geq t\mathbf{1}_{X \geq t}$ (Why?). Therefore $\mathbf{E}[Z] \geq t\mathbf{E}[\mathbf{1}_{X \geq t}] = t\mathbf{P}\{X \geq t\}$. Putting these together we get $\mathbf{E}[X] \geq t\mathbf{P}\{X \geq t\}$ as we desired to show. ■

Markov's inequality is simple, but surprisingly useful. Firstly, one can apply it to functions of our random variable and get many inequalities. Here are some.

Variants of Markov's inequality:

(1) If X is a non-negative random variable with finite p^{th} moment, then $\mathbf{P}\{X \geq t\} \leq t^{-p}\mathbf{E}[X^p]$ for any $t > 0$.

(2) If X is a random variable with finite second moment and $\mu = \mathbf{E}[X]$, then $\mathbf{P}[|X - \mu| \geq t] \leq \frac{1}{t^2}\text{Var}(X)$. [*Chebyshev's inequality*]

(3) If X is a random variable with finite exponential moments, then $\mathbf{P}\{X \geq t\} \leq e^{-\lambda t}\mathbf{E}[e^{\lambda X}]$ for any $\lambda > 0$. [*Chernoff's inequality*]

Thus, if we only know that X has finite mean, the tail probability $\mathbf{P}\{X \geq t\}$ must decay at least as fast as $1/t$. But, if we knew that the second moment was finite we could assert that the decay must be at least as fast as $1/t^2$, which is better. If $\mathbf{E}[e^{\lambda X}] < \infty$, then we get much faster decay of the tail, like $e^{-\lambda t}$.

Chebyshev's inequality captures again the intuitive notion that variance measures the spread of the distribution about the mean. The smaller the variance, lesser the spread. An alternate way to write Chebyshev's inequality is

$$\mathbf{P}(|X - \mu| > r\sigma) \leq \frac{1}{r^2},$$

where $\sigma = \text{s.d.}(X)$. This measures the deviations in multiples of the standard deviation. This is a very general inequality. In specific cases we can get better bounds than $1/r^2$ (just like Markov inequality can be improved using higher moments, when they exist).

Jensen's Inequality: Let $I \subseteq \mathbb{R}$ be an interval and let $\varphi : I \to \mathbb{R}$ be a twice differentiable function such that its second order derivative φ'' is continuous on I and $\varphi''(x) \geq 0 \, \forall \, x \in \mathbb{R}$ (i.e., φ is convex). Let X be a random variable with support $S_X \subseteq I$, and finite expectation. Then,

$$\mathbf{E}[\varphi(X)] \geq \varphi(\mathbf{E}[X]).$$

If $\varphi'(x) > 0 \, \forall \, x \in I$, then the inequality above is strict unless X is a degenerate random variable.

Exercise 17. A random variable X is said to be degenerate at a point $c \in \mathbb{R}$ if $\mathbf{P}(X = c) = 1$. Find $\text{Var}[X]$.

AM-GM-HM Inequality: Let X be a random variable with support $S_X \subseteq (0, \infty)$. Then, $\mathbf{E}[X]$ is called the arithmetic mean (AM) of X, $e^{\mathbf{E}[\ln X]}$ is called the geometric mean (GM) of X, and $\frac{1}{\mathbf{E}[1/X]}$ is called harmonic mean (HM) of X (provided these quantities exist). Then,

$$\mathbf{E}[X] \geq e^{\mathbf{E}[\ln X]} \geq \frac{1}{\mathbf{E}[1/X]}.$$

Exercise 18. Prove this inequality.

5. Simulation - I

As we have emphasized, probability is applicable to many situations in the real world. As such one may conduct experiments to verify the extent to which theorems are actually valid. For this we need to be able to draw numbers at random from any given distribution.

For example, take the case of Bernoulli($1/2$) distribution. One experiment that can give this is that of physically tossing a coin. This is not entirely satisfactory for several reasons. Firstly, are real coins fair? Secondly, what if we change slightly and want to generate from Ber(0.45)? In this section, we describe how to draw random numbers from various distributions on a computer. We do not fully answer this question. Instead what we shall show is

If one can generate random numbers from Unif($[0,1]$) distribution, then one can draw random numbers from any other distribution. More precisely, suppose U is a random variable with Unif($[0,1]$) distribution. We want to simulate random numbers from a given distribution F. Then, we shall find a function $\psi : [0,1] \to \mathbb{R}$ so that the random variable $X := \psi(U)$ has the given distribution F.

Important: The question of how to draw random numbers from Unif($[0,1]$) distribution is a very difficult one, and we shall just make a few superficial remarks about that.

Drawing random numbers from a discrete pmf: First start with an example.

Example 19. Suppose we want to draw random numbers from Ber(0.4) distribution. Let $\psi : [0,1] \to \mathbb{R}$ be defined as $\psi(t) = \mathbf{1}_{t \le 0.4}$. Let $X = \psi(U)$, i.e., $X = 1$ if $U \le 0.4$ and $X = 0$ otherwise. Then

$$\mathbf{P}\{X = 1\} = \mathbf{P}\{U \le 0.4\} = 0.4, \qquad \mathbf{P}\{X = 0\} = \mathbf{P}\{U > 0.4\} = 0.6.$$

Thus, X has Ber(0.4) distribution.

It is clear how to generalize this.

General rule: Suppose we are given a pmf f

$$\begin{pmatrix} t_1 & t_2 & t_3 & \dots \\ f(t_1) & f(t_2) & f(t_3) & \dots \end{pmatrix}.$$

Then, define $\psi : [0,1] \to \mathbb{R}$ as

$$\psi(u) = \begin{cases} t_1 & \text{if } u \in [0, f(t_1)] \\ t_2 & \text{if } u \in (f(t_1), f(t_1) + f(t_2)] \\ t_3 & \text{if } u \in (f(t_1) + f(t_2), f(t_1) + f(t_2) + f(t_3)] \\ \vdots & \vdots \end{cases}$$

Then, define $X = \psi(U)$. Clearly X takes the values $t_1, t_2, \ldots$ and

$$\mathbf{P}\{X = t_k\} = \mathbf{P}\left\{\sum_{j=1}^{k-1} f(t_j) < U \leq \sum_{j=1}^{k} f(t_j)\right\} = f(t_k).$$

Thus, X has pmf f.

Exercise 20. Write R codes to draw 100 random numbers from each of the following distributions (and draw the histograms). Compare with the pmf.

(1) Bin(n, p) for $n = 10, 20, 40$ and $p = 0.5, 0.3, 0.9$.

(2) Geo(p) for $p = 0.9, 0.5, 0.3$.

(3) Pois(λ) with $\lambda = 1, 4, 10$.

(4) Hypergeo(N_1, N_2, m) with $N_1 = 100, N_2 = 50, m = 20$, $N_1 = 1000, N_2 = 1000, m = 40$.

Check the file 'Distributions_discrete.R' under 'Resources' on HelloIITK.

1. Simulation - II

Drawing random numbers from a pdf: Clearly, the procedure used for generating from a pmf is inapplicable here. First start with two examples. As before, U is a Unif($[0,1]$) random variable.

Example 1. Suppose we want to draw from the Unif($[3,7]$) distribution. Set $X = 4U + 3$. Clearly,

$$F(t) := \mathbf{P}\{X \le t\} = \mathbf{P}\{U \le \frac{t-3}{4}\} = \begin{cases} 0 & \text{if } t < 3 \\ (t-3)/4 & \text{if } 3 \le t \le 7 \\ 1 & \text{if } t > 7. \end{cases}$$

This is precisely the CDF of Unif($[3,7]$) distribution.

Example 2. Here, let us do the opposite, just take some function of a uniform variable and see what CDF we get. Let $\psi(t) = t^3$ and let $X = \varphi(U) = U^3$. Then,

$$F(t) := \mathbf{P}\{X \le t\} = \mathbf{P}\{U \le t^{1/3}\} = \begin{cases} 0 & \text{if } t < 0 \\ t^{1/3} & \text{if } 0 \le t \le 1 \\ 1 & \text{if } t > 1. \end{cases}$$

Differentiating the CDF, we get the density

$$f(t) = F'(t) = \begin{cases} \frac{1}{3}t^{-2/3} & \text{if } 0 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The derivative does not exist at 0 and 1, but as remarked earlier, it does not matter if we change the value of the density at finitely many points (as the integral over any interval will remain the same). Anyway, we notice that the density is that of Beta($1/3, 1$). Hence, $X \sim$ Beta($1/3, 1$).

This gives us the idea that to generate random number from a continuous CDF F, we should find a function $\psi : [0,1] \to \mathbb{R}$ such that $X := \psi(U)$ has the distribution F. How to find the distribution of X?

Lemma 3. *Let $\psi : (0,1) \to \mathbb{R}$ be a strictly increasing function with $a = \psi(0+)$ and $b = \psi(1-)$. Let $X = \psi(U)$ with $U \sim U(0,1)$. Then, X has CDF*

$$F(t) = \begin{cases} 0 & \text{if } t \le a \\ \psi^{-1}(t) & \text{if } a < t < b \\ 1 & \text{if } t \ge b. \end{cases}$$

If ψ is also differentiable and the derivative does not vanish anywhere (or, vanishes at finitely many points only), then X has pdf

$$f(t) = \begin{cases} \left(\psi^{-1}\right)'(t) & \textit{if } a < t < b \\ 0 & \textit{if } t \notin (a,b). \end{cases}$$

Proof. Since ψ is strictly increasing, $\psi(u) \le t$ if and only if $u \le \psi^{-1}(t)$. Hence,

$$F(t) = \mathbf{P}\{X \le t\} = \mathbf{P}\{U \le \psi^{-1}(t)\} = \begin{cases} 0 & \text{if } t \le a \\ \psi^{-1}(t) & \text{if } a < t < b \\ 1 & \text{if } t \ge b. \end{cases}$$

■

From this lemma, we immediately get the following rule for generating random numbers from a density.

How to simulate from a CDF: Let F be a CDF that is strictly increasing on an interval $[A,B]$, where $F(A) = 0$ and $F(B) = 1$ (it is allowed to take $A = -\infty$ and/or $B = +\infty$). Then, define $\psi : (0,1) \to (A,B)$ as $\psi(u) = F^{-1}(u)$. Let $U \sim \text{Unif}([0,1])$ and let $X = \psi(U)$. Then, X has CDF equal to F.

This follows from the lemma because ψ is defined as the inverse of F, and hence F (restricted to (A,B)) is the inverse of ψ. Further, as the inverse of a strictly increasing function, the function ψ is also strictly increasing.

Example 4. Consider the Exponential distribution with parameter λ whose CDF is

$$F(t) = \begin{cases} 0 & \text{if } t \le 0 \\ 1 - e^{-\lambda t} & \text{if } t > 0. \end{cases}$$

Take $A = 0$ and $B = +\infty$. Then F is increasing on $(0,\infty)$ and its inverse is the function $\psi(u) = -\frac{1}{\lambda}\log(1-u)$. Thus to simulate a random number from Exp(λ) distribution, we set $X = -\frac{1}{\lambda}\log(1-U)$.

When the CDF is NOT explicitly available as a function, we can still adopt the above procedure, but only numerically. Consider an example.

Example 5. Suppose $F = \Phi$, the CDF of $N(0,1)$ distribution. Then, we do not have an explicit form for either Φ or for its inverse Φ^{-1}. With a computer we can do the following. Pick a large number of closely placed points, for example divide the interval $[-5,5]$ into 1000 equal intervals of length 0.01 each. Let the endpoints of these intervals be labelled $t_0 < t_1 < \cdots < t_{1000}$. For each

i, calculate $\Phi(t_i) = \int_{-\infty}^{t_i} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ using numerical methods for integration, say the numerical value obtained is w_i. This is done only once, and create the table of values

$$\begin{matrix} t_0 & t_1 & t_2 & \dots & \dots & t_{1000} \\ w_0 & w_1 & w_2 & \dots & \dots & w_{1000} \end{matrix}.$$

Now, draw a uniform random number U. Look up the table and find the value of i for which $w_i < U < w_{i+1}$. Then set $X = t_i$. If it so happens that $U < w_0$, set $X = t_0 = -5$ and if $U > w_{1000}$ set $X = t_{1000} = 5$. But since $\Phi(-5) < 0.00001$ and $\Phi(5) > 0.99999$, it is highly unlikely that the last two cases will occur. The random variable X has a distribution close to $N(0, 1)$.

Exercise 6. Give an explicit method to draw random numbers from the following densities.

(1) Cauchy distribution with density $\frac{1}{\pi(1+x^2)}$.

(2) Beta($\frac{1}{2}, \frac{1}{2}$) density $\frac{1}{\pi} \frac{1}{\sqrt{x(1-x)}}$ on $[0, 1]$ (and zero elsewhere).

(3) Pareto(α) distribution which by definition has the density

$$f(t) = \begin{cases} \alpha t^{-\alpha-1} & \text{if } t \geq 1, \\ 0 & \text{if } t < 1. \end{cases}$$

Check the file 'Distributions_continuous.R'.

Remark 7. We have conveniently skipped the question of how to draw random numbers from the uniform distribution in the first place. This is a difficult topic and various results, proved and unproved, are used in generating such numbers.

2. Joint distributions

In many situations we study several random variables at once. In such a case, knowing the individual distributions is not sufficient to answer all relevant questions. This is like saying that knowing $\mathbf{P}(A)$ and $\mathbf{P}(B)$ is insufficient to calculate $\mathbf{P}(A \cap B)$ or $\mathbf{P}(A \cup B)$ etc.

Definition 8 (Joint distribution)**.** Let $X_1, X_2, \ldots, X_m$ be random variables on the same probability space. We call $\mathbf{X} = (X_1, \ldots, X_m)$ a *random vector*, as it is just a vector of random variables. The CDF of $\mathbf{X}$, also called the joint CDF of $X_1, \ldots, X_m$ is the function $F_{\mathbf{X}} : \mathbb{R}^m \to [0,1]$ defined as

$$F_{\mathbf{X}}(t_1, \ldots, t_m) = \mathbf{P}\{X_1 \le t_1, \ldots, X_m \le t_m\} = \mathbf{P}\left\{\bigcap_{i=1}^{m} \{X_i \le t_i\}\right\}.$$

Exercise 9. Consider two events A and B in the probability space and let $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$ be their indicator random variables. Find their joint CDF.

Properties of joint CDFs: The following properties of the joint CDF $F_{\mathbf{X}} : \mathbb{R}^m \to [0,1]$ are analogous to those of the 1-dimensional CDF and the proofs are similar.

1. $F_{\mathbf{X}}$ is increasing in each co-ordinate, i.e., if $s_1 \le t_1, \ldots, s_m \le t_m$, then $F_{\mathbf{X}}(s_1, \ldots, s_m) \le F_{\mathbf{X}}(t_1, \ldots, t_m)$.
2. $\lim F_{\mathbf{X}}(t_1, \ldots, t_m) = 0$ if $\max\{t_1, \ldots, t_m\} \to -\infty$ (i.e., one of the t_i goes to $-\infty$).
3. $\lim F_{\mathbf{X}}(t_1, \ldots, t_m) = 1$ if $\min\{t_1, \ldots, t_m\} \to +\infty$ (i.e., all of the t_i goes to $+\infty$).
4. $F_{\mathbf{X}}$ is right continuous in each co-ordinate. That is $F_{\mathbf{X}}(t_1+h_1, \ldots, t_m+h_m) \to F_{\mathbf{X}}(t_1, \ldots, t_m)$ as $h_i \to 0^+$ for $i = 1, \ldots, m$.

Conversely, any function having these four properties is the joint CDF of some random variable(s).

Remark 10. Recall that the increasing property in $\mathbb{R}$ is $F(t_1) - F(s_1) = \mathbf{P}(s_1 < X_1 \le t_1) \ge 0$ for $s_1 \le t_1$. We draw a similar analogy to generalize this to $\mathbb{R}^2$, and need to verify the following:

$$\mathbf{P}(s_1 < X_1 \le t_1, s_2 < X_2 \le t_2) = F(t_1, t_2) - F(s_1, t_2) - F(t_1, s_2) + F(s_1, s_2) \ge 0$$

for $s_1 \le t_1, s_2 \le t_2$. A general expression can be derived for $\mathbb{R}^m$, but it is more complicated.

From the joint CDF, it is easy to recover the individual CDFs. Indeed, if $F_{\mathbf{X}} : \mathbb{R}^m \to [0,1]$ is the CDF of $\mathbf{X} = (X_1, \ldots, X_m)$, then the CDF of X_1 is given by $F_1(t) := F_{\mathbf{X}}(t, +\infty, \ldots, +\infty) := \lim F_{\mathbf{X}}(t, s_2, \ldots, s_m)$ as $s_i \to +\infty$ for each $i = 2, \ldots, m$. This is true because if $A_n := \{X_1 \le t\} \cap \{X_2 \le n\} \cap \cdots \cap \{X_m \le n\}$, then the events A_n increase to the event $A = \{X_1 \le t\}$ as $n \to \infty$. Hence, $\mathbf{P}(A_n) \to \mathbf{P}(A)$. But, $\mathbf{P}(A_n) = F_{\mathbf{X}}(t, n, \ldots, n)$ and $\mathbf{P}(A) = F_1(t)$. Thus, we see that $F_1(t) := F_{\mathbf{X}}(t, +\infty, \ldots, +\infty)$.

More generally, we can recover the joint CDF of any subset of $X_1, \dots, X_n$. For example, the joint CDF of $X_1, \dots, X_k$ is just $F_{\mathbf{X}}(t_1, \dots, t_k, +\infty, \dots, +\infty)$.

Joint pmf and pdf: Just like in the case of one random variable, we can consider the following two sub-classes of vector of random variables.

(1) Distributions with a pmf. These are CDFs for which there exist points $\mathbf{t}_1, \mathbf{t}_2, \dots$ in $\mathbb{R}^m$ and non-negative numbers w_i such that $\sum_i w_i = 1$ (often we write $f(\mathbf{t}_i)$ in place of w_i). For every $\mathbf{t} \in \mathbb{R}^m$, we have

$$F(\mathbf{t}) = \sum_{i\,:\,\mathbf{t}_i \leq \mathbf{t}} w_i,$$

where $\mathbf{s} \leq \mathbf{t}$ means that each co-ordinate of $\mathbf{s}$ is less than, or equal to the corresponding co-ordinate of $\mathbf{t}$.

(2) Distributions with a pdf. These are CDFs for which there is a non-negative function (may assume piecewise continuous for convenience) $f : \mathbb{R}^m \to \mathbb{R}_+$ such that for every $\mathbf{t} \in \mathbb{R}^m$ we have

$$F(\mathbf{t}) = \int\limits_{-\infty}^{t_1} \dots \int\limits_{-\infty}^{t_m} f(u_1, \dots, u_m)\, du_1 \dots du_m.$$

We give two examples, one of each kind.

Example 11. (Multinomial distribution). Fix parameters r, m (two positive integers) and $p_1, \dots, p_m$ (positive numbers that add to 1). The *multinomial pmf* with these parameters is given by

$$f(k_1, \dots, k_{m-1}) = \frac{r!}{k_1! k_2! \cdots k_{m-1}!(r - \sum_{i=1}^{m-1} k_i)!} p_1^{k_1} \dots p_{m-1}^{k_{m-1}} p_m^{r - \sum_{i=1}^{m-1} k_i},$$

if $k_i \geq 0$ are integers such that $k_1 + \dots + k_{m-1} \leq r$. [1] One situation where this distribution arises is when r balls are randomly placed in m bins, with each ball going into the jth bin with probability p_j, and we look at the random vector $(X_1, \dots, X_{m-1})$, where X_k is the number of balls that fell into the kth bin. This random vector has the multinomial pmf.

In this case, the marginal distribution of X_k is Bin(r, p_k). More generally, $(X_1, \dots, X_\ell)$ has multinomial distribution with parameters $r, \ell, p_1, \dots, p_\ell, p_0$, where $p_0 = 1 - (p_1 + \dots + p_\ell)$. This is easy to prove, but even easier to see from the balls in bins interpretation (just think of the last $n - \ell$ bins as one).

[1] In some books, the distribution of $(X_1, \dots, X_m)$ is called the multinomial distribution. This has the pmf

$$g(k_1, \dots, k_m) = \frac{r!}{k_1! k_2! \cdots k_{m-1}! k_m!} p_1^{k_1} \cdots p_{m-1}^{k_{m-1}} p_m^{k_m},$$

where k_i are non-negative integers such that $k_1 + \dots + k_m = r$. We have chosen our convention so that the binomial distribution is a special case of the multinomial.

Example 12. (Bivariate normal distribution). Consider a density function on $\mathbb{R}^2$ given by

$$f(x,y) = \frac{\sqrt{ab-c^2}}{2\pi} e^{-\frac{1}{2}\left[a(x-\mu)^2+b(y-\nu)^2+2c(x-\mu)(y-\nu)\right]},$$

where μ, ν, a, b, c are real parameters. We shall impose the conditions that $a > 0$, $b > 0$ and $ab - c^2 > 0$ (otherwise the above does not give a density, as we shall see).

The first thing is to check that this is indeed a density. We recall the one-dimensional Gaussian integral

$$(1) \qquad \int_{-\infty}^{+\infty} e^{-\frac{\tau}{2}(x-a)^2} dx = \sqrt{2\pi}\frac{1}{\sqrt{\tau}} \text{ for any } \tau > 0 \text{ and any } a \in \mathbb{R}.$$

We shall take $\mu = \nu = 0$ (how do you compute the integral if they are not?). Then, the exponent in the density has the form

$$ax^2 + by^2 + 2cxy = b\left(y + \frac{cx}{b}\right)^2 + \left(a - \frac{c^2}{b}\right)x^2.$$

Therefore,

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[ax^2+by^2+2cxy\right]} dy &= e^{-\frac{1}{2}(a-\frac{c^2}{b})x^2} \int_{-\infty}^{\infty} e^{-\frac{b}{2}(y+\frac{cx}{b})^2} dy \\ &= e^{-\frac{1}{2}(a-\frac{c^2}{b})x^2} \frac{\sqrt{2\pi}}{\sqrt{b}}\end{aligned}$$

by (1) but ony if $b > 0$. Now, we integrate over x and use (1) again (also the fact that $a - \frac{c^2}{b} > 0$) to get

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[a(x-\mu)^2+b(y-\nu)^2+2c(x-\mu)(y-\nu)\right]} dy dx &= \frac{\sqrt{2\pi}}{\sqrt{b}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(a-\frac{c^2}{b})x^2} dx \\ &= \frac{\sqrt{2\pi}}{\sqrt{b}} \frac{\sqrt{2\pi}}{\sqrt{a-\frac{c^2}{b}}} = \frac{2\pi}{\sqrt{ab-c^2}}.\end{aligned}$$

This completes the proof that $f(x,y)$ is indeed a density. Note that $b > 0$ and $ab - c^2 > 0$ also implies that $a > 0$.

Matrix form of writing the density: Let $\Sigma^{-1} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$. Then, $\det(\Sigma) = \frac{1}{\det(\Sigma^{-1})} = \frac{1}{ab-c^2}$. Hence, we may re-write the density above as (let $\mathbf{u}$ be the column vector with co-ordinates x, y)

$$f(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}\mathbf{u}^t \Sigma^{-1}\mathbf{u}}.$$

The conditions $a > 0, b > 0, ab - c^2 > 0$ translate precisely to what is called positive-definiteness. One way to say it is that Σ is a symmetric matrix and all its eigenvalues are strictly positive.

Final form: We can now introduce an extra pair of parameters μ_1, μ_2 and define a density

$$f(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{u}-\mu)^t \Sigma^{-1}(\mathbf{u}-\mu)},$$

where μ is a column vector with co-ordinates μ_1, μ_2. This is the full bi-variate normal density. Along similar lines, we can talk of the m-variate normal density.

Example 13. (A class of examples). Let $f_1, f_2, \ldots, f_m$ be one-variable densities. In other words, $f_i : \mathbb{R} \to \mathbb{R}_+$ and $\int_{-\infty}^{\infty} f_i(x)dx = 1$. Then, we can make a multivariate density as follows. Define $f : \mathbb{R}^m \to \mathbb{R}_+$ by

$$f(x_1, \ldots, x_m) = f_1(x_1) \cdots f_m(x_m).$$

Then, f is a density.

If X_i are random variables on a common probability space and the joint density of $\mathbf{X} = (X_1, \ldots, X_m)$ is $f(x_1, \ldots, x_m)$, then we say that X_i are *independent random variables*. It is easy to see that the marginal density of X_i is f_i for $i = 1, \ldots, m$. It is also the case that the joint CDF factors as

$$F_{\mathbf{X}}(x_1, \ldots, x_m) = F_{X_1}(x_1) \cdots F_{X_m}(x_m).$$

Example 14. (A second class of examples). Let g be a pdf. In other words, $g : \mathbb{R} \to \mathbb{R}_+$ and $\int_{-\infty}^{\infty} g(x)dx = 1$. Then, we can make a multivariate density as follows. Define $f : \mathbb{R}^m \to \mathbb{R}_+$ by

$$f(x_1, \ldots, x_m) = g(x_1) \cdots g(x_m).$$

Then, f is a density.

If X_i are random variables on a common probability space and the joint density of $\mathbf{X} = (X_1, \ldots, X_m)$ is $f(x_1, \ldots, x_m)$, then we say that X_i are *independent and identically distributed (i.i.d.) random variables*. It is easy to see that the marginal density of X_i is g for $i = 1, \ldots, m$. It is also the case that the joint CDF factors as

$$F_{\mathbf{X}}(x_1, \ldots, x_m) = G(x_1) \cdots G(x_m).$$

3. CHANGE OF VARIABLE IN $\mathbb{R}^m$

CDF technique: Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a random vector and let g be a function such that $g : \mathbb{R}^m \to \mathbb{R}$. The distribution $Y = g(X_1, \ldots, X_m)$ can be determined by computing the distribution function

$$F_Y(y) = \mathbf{P}\left(\{g(X_1, \ldots, X_m) \leq y\}\right), \quad -\infty < y < \infty.$$

Example 15. Let X_1, X_2 be i.i.d. from the $U(0,1)$ distribution. Find the distribution function of $Y = X_1 + X_2$. Hence, find the pdf of Y.

The joint pdf of (X_1, X_2) is given by

$$\begin{aligned} f_{X_1,X_2}(x_1, x_2) &= f_{X_1}(x_1) f_{X_2}(x_2) \\ &= \begin{cases} 1, & \text{if } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Therefore, the distribution function of Y is given by

$$\begin{aligned} F_Y(x) &= \mathbf{P}\left(\{X_1 + X_2 \leq x\}\right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2) \mathbf{1}_{(-\infty,x]}(x_1 + x_2) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \\ &= \int_0^1 \int_0^1 \mathbf{1}_{(0,x]}(x_1 + x_2) \; \mathrm{d}x_1 \, \mathrm{d}x_2 \\ &= \begin{cases} 0, & \text{if } x < 0 \\ \frac{1}{2} \times x \times x, & \text{if } 0 \leq x < 1 \\ 1 - \frac{1}{2} \times (2 - x) \times (2 - x), & \text{if } 1 \leq x < 2 \\ 1, & \text{if } x \geq 2 \end{cases} \\ &= \begin{cases} 0, & \text{if } x < 0 \\ \frac{x^2}{2}, & \text{if } 0 \leq x < 1 \\ \frac{4x - x^2 - 2}{2}, & \text{if } 1 \leq x < 2 \\ 1, & \text{if } x \geq 2. \end{cases} \end{aligned}$$

Clearly, F_Y is differentiable everywhere except on a finite set $\{0, 1, 2\}$. Let

$$g(x) = \begin{cases} x, & \text{if } 0 < x < 1 \\ 2 - x, & \text{if } 1 < x < 2 \\ 0, & \text{otherwise,} \end{cases}$$

so that

$$\frac{\mathrm{d}}{\mathrm{d}x}F_Y(x) = g(x)\;\forall\, x \in \mathbb{R}\setminus\{0,1,2\} \text{ and } \int_{-\infty}^{\infty} g(x)\mathrm{d}x = 1.$$

Change of variable for discrete distributions: The idea is quite similar to the univariate case, and will be skipped. One example is discussed.

Example 16. Let $\mathbf{X} = (X_1, X_2, X_3)$ be a discrete type random vector with pmf

$$f_{\mathbf{X}}(x_1,x_2,x_3) = \begin{cases} \frac{2}{9}, & \text{if } (x_1,x_2,x_3) \in \{(1,1,0),(1,0,1),(0,1,1)\} \\ \frac{1}{3}, & \text{if } (x_1,x_2,x_3) = (1,1,1) \\ 0, & \text{otherwise.} \end{cases}$$

Define $Y_1 = X_1 + X_2$ and $Y_2 = X_2 + X_3$. Find the marginal pmf of Y_1 and Y_2.

We have

$$\mathbf{P}(\{Y_1 = y\}) = \mathbf{P}(\{X_1 + X_2 = y\}) = 0, \quad \text{if } y \notin \{1,2\}.$$

$$\begin{aligned} \mathbf{P}(\{Y_1 = 1\}) &= \mathbf{P}(\{X_1 + X_2 = 1\}) \\ &= \mathbf{P}(\{(X_1,X_2,X_3) \in \{(1,0,1),(0,1,1)\}\}) \\ &= \mathbf{P}(\{(X_1,X_2,X_3) = (1,0,1)\}) + \mathbf{P}(\{(X_1,X_2,X_3) = (0,1,1)\}) \\ &= \tfrac{4}{9}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{P}(\{Y_1 = 2\}) &= \mathbf{P}(\{X_1 + X_2 = 2\}) \\ &= \mathbf{P}(\{(X_1,X_2,X_3) = (1,1,0)\}) + \mathbf{P}(\{(X_1,X_2,X_3) = (1,1,1)\}) \\ &= \tfrac{5}{9}. \end{aligned}$$

Therefore,

$$f_{Y_1}(y) = \begin{cases} \frac{4}{9}, & \text{if } y = 1 \\ \frac{5}{9}, & \text{if } y = 2 \\ 0, & \text{otherwise.} \end{cases}$$

By symmetry of $f_{\mathbf{X}}$, we get

$$f_{Y_2}(y) = \begin{cases} \frac{4}{9}, & \text{if } y = 1 \\ \frac{5}{9}, & \text{if } y = 2 \\ 0, & \text{otherwise.} \end{cases}$$

Exercise 17. Find the joint pmf of $\underline{Y} = (Y_1, Y_2)$. Are Y_1 and Y_2 independent?

Change of variable for continuous distributions: The idea is quite similar to the univariate case, and will be discussed for a *special* class of functions.

Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a random vector with density $f(t_1, \ldots, t_m)$. Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be a *one-one function which is continuously differentiable* (some exceptions can be made, as remarked later).

Let $\mathbf{Y} = T(\mathbf{X})$. In co-ordinates, we may write $\mathbf{Y} = (Y_1, \ldots, Y_m)$ and $Y_1 = T_1(X_1, \ldots, X_m), \ldots,$ $Y_m = T_m(X_1, \ldots, X_m)$, where $T_i : \mathbb{R}^m \to \mathbb{R}$ are the components of T.

Question: What is the joint density of $Y_1, \ldots, Y_m$?

The change of variable formula: In the setting described above, the joint density of $Y_1, \ldots, Y_m$ is given by

$$g(\mathbf{y}) = f\left(T^{-1}\mathbf{y}\right) \, \big| \, J[T^{-1}](\mathbf{y}) \, \big| \, ,$$

where $|J[T^{-1}](\mathbf{y})|$ is the Jacobian determinant of the function T^{-1} at the point $\mathbf{y} = (y_1, \ldots, y_m)$.

Enlarging the applicability of the change of variable formula: The change of variable formula is applicable in greater generality than we stated above.

(1) Firstly, T does not have to be defined on all of $\mathbb{R}^m$. It is sufficient if it is defined on the range of $\mathbf{X}$ (i.e., if $f(t_1, \ldots, t_m) = 0$ for $(t_1, \ldots, t_m) = \mathbb{R}^m \setminus A$), then it is enough if T is defined on A.

(2) Similarly, the differentiability of T is required only on a subset, outside of which $\mathbf{X}$ has probability 0 of falling. For example, finitely many points, on a line (if $m \geq 2$), or on a plane (if $m \geq 3$), etc.

(3) One-one property of T is important, but there are special cases which can be dealt with by a slight modification. For example, if $T(x) = x^2$ or $T(x_1, x_2) = (x_1^2, x_2^2)$, where we can split the space into parts on each of which T is one-one.

Example 18. Let X_1, X_2 be independent Exp(λ) random variables. Let $T(x_1, x_2) = (x_1 + x_2, \frac{x_1}{x_1 + x_2})$. This is well-defined on $\mathbb{R}_+^2$ (and note that $\mathbf{P}\{(X_1, X_2) \in \mathbb{R}_+^2\} = 1$) and its range is $\mathbb{R}_+ \times (0, 1)$. The inverse function is $T^{-1}(y_1, y_2) = (y_1 y_2, y_1(1 - y_2))$. Its Jacobian determinant is

$$|J[T^{-1}](y_1, y_2)| = \det \begin{bmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{bmatrix} = -y_1.$$

(X_1, X_2) has density $f(x_1, x_2) = \lambda^2 e^{-\lambda(x_1 + x_2)}$ for $x_1, x_2 > 0$ (henceforth, if not mentioned explicitly, it will be a convention that the density is zero except where we specify it). Hence, the random variables $Y_1 = X_1 + X_2$ and $Y_2 = \frac{X_1}{X_1 + X_2}$ have joint density

$$g(y_1, y_2) = f(y_1 y_2, y_1(1 - y_2))|J[T^{-1}](y_1, y_2)| = \lambda^2 e^{-\lambda(y_1 y_2 + y_1(1 - y_2))} y_1 = \lambda^2 y_1 e^{-\lambda y_1}$$

for $y_1 > 0$ and $y_2 \in (0,1)$.

In particular, we see that $Y_1 = X_1 + X_2$ has density $h_1(t) = \int_0^1 \lambda^2 te^{-\lambda t} ds = \lambda^2 te^{-\lambda t}$ (for $t > 0$) which means that $Y_1 \sim \text{Gamma}(2,\lambda)$. Similarly, $Y_2 = \frac{X_1}{X_1+X_2}$ has density $h_2(s) = \int_0^\infty \lambda^2 te^{-\lambda t} dt = 1$ (for $s \in (0,1)$) which means that Y_2 has $\text{Unif}(0,1)$ distribution. In fact, Y_1 and Y_2 are also independent since $g(u,v) = h_1(u)h_2(v)$.

Exercise 19. Let $X_1 \sim \text{Gamma}(\nu_1,\lambda)$ and $X_2 \sim \text{Gamma}(\nu_2,\lambda)$ (note that the shape parameter is the same) and assume that they are independent. Find the joint distribution of X_1+X_2 and $\frac{X_1}{X_1+X_2}$.

Example 20. Suppose we are given that X_1 and X_2 are independent and each has $\text{Exp}(\lambda)$ distribution. What is the distribution of the random variable $X_1 + X_2$?

The change of variable formula works for transformations from $\mathbb{R}^m$ to $\mathbb{R}^m$ whereas here we have two random variables X_1, X_2 and our interest is in one random variable $X_1 + X_2$. To use the change of variable formula, we must introduce an *auxiliary* variable. For example, we take $Y_1 = X_1 + X_2$ and $Y_2 = X_1/(X_1 + X_2)$. Then as in the first example, we find the joint density of (Y_1, Y_2) using the change of variable formula and then integrate out the second variable to get the density of Y_1.

Let us emphasize the point that if our interest is only in Y_1, then we have a lot of freedom in choosing the auxiliary variable. The only condition is that from Y_1 and Y_2 we should be able to recover X_1 and X_2. Let us repeat the same using $Y_1 = X_1 + X_2$ and $Y_2 = X_2$. Then, $T(x_1,x_2) = (x_1+x_2, x_2)$ maps $\mathbb{R}_+^2$ onto $Q := \{(y_1,y_2) : y_1 > y_2 > 0\}$ in a one-one manner. The inverse function is $T^{-1}(y_1,y_2) = (y_1-y_2, y_2)$. It is easy to see that $|J[T^{-1}](y_1,y_2)| = 1$ (check!). Hence, by the change of variable formula, the density of (Y_1,Y_2) is given by

$$\begin{aligned} g(y_1,y_2) &= f(y_1-y_2,y_2)\cdot 1 \\ &= \lambda^2 e^{-\lambda(y_1-y_2)}e^{-\lambda y_2} \quad (\text{if } y_1 > y_2 > 0) \\ &= \lambda^2 e^{-\lambda y_1}\mathbf{1}_{y_1>y_2>0}. \end{aligned}$$

To get the density of Y_1, we integrate out the second variable. The density of Y_1 is

$$\begin{aligned} h(u) &= \int\limits_{-\infty}^{\infty} \lambda^2 e^{-\lambda y_1}\mathbf{1}_{y_1>y_2>0} dy_2 \\ &= \lambda^2 e^{-\lambda y_1}\int\limits_0^{y_1} dy_2 \\ &= \lambda^2 y_1 e^{-\lambda y_1} \end{aligned}$$

which agrees with what we found before.

Example 21. Suppose $R \sim \text{Exp}(\lambda)$ and $\Theta \sim \text{Unif}(0, 2\pi)$ and the two are independent. Define $X = \sqrt{R}\cos(\Theta)$ and $Y = \sqrt{R}\sin(\Theta)$. We want to find the distribution of (X, Y). For this, we first write the joint density of (R, Θ) which is given by

$$f(r, \theta) = \frac{1}{2\pi}\lambda e^{-\lambda r} \quad \text{for } r > 0, \theta \in (0, 2\pi).$$

Define the transformation $T : \mathbb{R}_+ \times (0, 2\pi) \to \mathbb{R}^2$ by $T(r, \theta) = (\sqrt{r}\cos\theta, \sqrt{r}\sin\theta)$. The image of T consists of all $(x, y) \in \mathbb{R}^2$ with $y \neq 0$. The inverse is $T^{-1}(x, y) = (x^2 + y^2, \arctan(y/x))$, where $\arctan(y/x)$ is defined so as to take values in $(0, \pi)$ when $y > 0$ and to take values in $(\pi, 2\pi)$ when $y < 0$. Thus

$$|J[T^{-1}](x, y)| = \det \begin{bmatrix} 2x & 2y \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix} = 2.$$

Therefore, (X, Y) has joint density

$$g(x, y) = 2f(x^2 + y^2, \arctan(y/x)) = \frac{\lambda}{\pi} e^{-\lambda(x^2+y^2)}.$$

This is for $(x, y) \in \mathbb{R}^2$ with $y \neq 0$. Since $g(x, y)$ separates into a function of x and a function of y, X, Y are independent $N(0, \frac{1}{2\lambda})$.

Remark 22. Relationships between random variables derived by the change of variable formulas can be used for simulation too. For instance, the CDF of $N(0, 1)$ is not explicit and hence simulating from that distribution is difficult (must resort to numerical methods). However, we can easily simulate it as follows. Simulate an $\text{Exp}(1/2)$ random variable R (easy, as the distribution function can be inverted) and simulate an independent $\text{Unif}(0, 2\pi)$ random variable Θ. Then, set $X = \sqrt{R}\cos(\Theta)$ and $Y = \sqrt{R}\sin(\Theta)$. These are two independent $N(0, 1)$ random numbers. Here, it should be noted that the random numbers in $(0, 1)$ given by a random number generator are supposed to be independent uniform random numbers.

4. Independence of random variables

Definition 23. Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a random vector (this means that X_i are random variables on a common probability space). We say that X_is are *independent* if

$$F_{\mathbf{X}}(t_1, \ldots, t_m) = F_1(t_1) \cdots F_m(t_m) \text{ for all } t_1, \ldots, t_m.$$

Remark 24. Recalling the definition of independence of events, the equality $F_{\mathbf{X}}(t_1, \ldots, t_m) = F_1(t_1) \cdots F_m(t_m)$ is just saying that the events $\{X_1 \le t_1\}, \ldots, \{X_m \le t_m\}$ are independent. More generally, it is true that $X_1, \ldots, X_m$ are independent if and only if $\{X_1 \in A_1\}, \ldots, \{X_m \in A_m\}$ are independent events for any $A_1, \ldots, A_m \subseteq \mathbb{R}$.

Remark 25. In case $X_1, \ldots, X_m$ have a joint pmf or a joint pdf (which we denote by $f(t_1, \ldots, t_m)$), the condition for independence is equivalent to

$$f(t_1, \ldots, t_m) = f_1(t_1) \cdots f_m(t_m),$$

where f_i is the marginal density (or pmf) of X_i. This fact can be derived from the definition easily. For example, in the case of densities, observe that

$$\begin{aligned} f(t_1, \ldots, t_m) &= \frac{\partial^m}{\partial t_1 \ldots \partial t_m} F(t_1, \ldots, t_m) \quad \text{(true for any joint density)} \\ &= \frac{\partial^m}{\partial t_1 \ldots \partial t_m} F_1(t_1) \cdots F_m(t_m) \quad \text{(by independence)} \\ &= F_1'(t_1) \cdots F_m'(t_m) \\ &= f_1(t_1) \cdots f_m(t_m). \end{aligned}$$

When we turn it around, this gives us a quicker way to check independence.

Fact: Let $X_1, \ldots, X_m$ be random variables with joint pdf $f(t_1, \ldots, t_m)$. Suppose we can write this pdf as

$$f(t_1, \ldots, t_m) = c g_1(t_1) \cdots g_m(t_m),$$

where c is a constant and g_i are some functions of one-variable. Then, $X_1, \ldots, X_m$ are independent. Further, the marginal density of X_k is $c_k g_k(t)$, where $c_k = \frac{1}{\int_{-\infty}^{+\infty} g_k(s) ds}$. An analogous statement holds when $X_1, \ldots, X_m$ have a joint pmf instead of pdf.

Example 26. Let $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = p^{\sum \omega_k} q^{n - \sum \omega_k}$. Define $X_k : \Omega \to \mathbb{R}$ by $X_k(\underline{\omega}) = \omega_k$. In words, we are considering the probability space corresponding to n tosses of a fair coin and X_k is the result of the kth toss. We claim that $X_1, \ldots, X_n$ are independent. Indeed, the joint pmf of $X_1, \ldots, X_n$ is

$$f(t_1, \ldots, t_n) = p^{\sum t_k} q^{n - \sum t_k}, \quad \text{where } t_i = 0 \text{ or } 1 \text{ for each } i \le n.$$

Clearly, $f(t_1,\dots,t_n) = g(t_1)g(t_2)\cdots g(t_n)$, where $g(s) = p^s q^{1-s}$ for $s = 0$ or 1 (this is just a terse way of saying that $g(s) = p$ if $s = 1$ and $g(s) = q$ if $s = 0$). Hence, $X_1,\dots,X_n$ are independent and X_k has pmf g (i.e., $X_k \sim \text{Ber}(p)$).

Example 27. Let (X, Y) have the bivariate normal density

$$f(x,y) = \frac{\sqrt{ab-c^2}}{2\pi} e^{-\frac{1}{2}(a(x-\mu_1)^2+b(y-\mu_2)^2+2c(x-\mu_1)(y-\mu_2))}.$$

If $c = 0$, we observe that

$$f(x,y) = C_0 e^{-\frac{a(x-\mu_1)^2}{2}} e^{-\frac{b(y-\mu_2)^2}{2}} \qquad (C_0 \text{ is a constant, exact value unimportant})$$

from which we deduce that X and Y are independent, and $X \sim N(\mu_1, \frac{1}{a})$ while $Y \sim N(\mu_2, \frac{1}{b})$.

Exercise 28. Can you argue that if $c \neq 0$, then X and Y are not independent?

Example 29. Let $\mathbf{X} = (X_1, X_2, X_3)$ be a random vector of absolutely continuous type with pdf

$$f_{\mathbf{X}}(x_1,x_2,x_3) = \begin{cases} \frac{1}{x_1 x_2}, & \text{if } 0 < x_3 < x_2 < x_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Are X_1, X_2 and X_3 independent random variables?

We have

$$f_{X_1}(x_1) = \begin{cases} \int_0^{x_1}\int_0^{x_2} \frac{1}{x_1 x_2}\,\mathrm{d}x_3\,\mathrm{d}x_2 = 1, & \text{if } 0 < x_1 < 1 \\ 0, & \text{other wise.} \end{cases}$$

$$f_{X_2}(x_2) = \begin{cases} \int_{x_2}^{1}\int_0^{x_2} \frac{1}{x_1 x_2}\,\mathrm{d}x_3\,\mathrm{d}x_1 = -\ln x_2, & \text{if } 0 < x_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

$$f_{X_3}(x_3) = \begin{cases} \int_{x_3}^{1}\int_{x_2}^{1} \frac{1}{x_1 x_2}\,\mathrm{d}x_1\,\mathrm{d}x_2 = \frac{(\ln x_3)^2}{2}, & \text{if } 0 < x_3 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

So,

$$f_{\mathbf{X}}(x_1,x_2,x_3) \neq f_{X_1}(x_1)\, f_{X_2}(x_2)\, f_{X_3}(x_3)\, \forall\, (x_1,x_2,x_3) \in \mathbb{R}^3,$$

and therefore X_1, X_2 and X_3 are not independent.

A very useful (and intuitively acceptable!) fact about independence is as follows.

Fact: Suppose $X_1,\dots,X_n$ are independent random variables. Let $k_1 < k_2 < \cdots < k_m = n$. Let $Y_1 = h_1(X_1,\dots,X_{k_1})$, $Y_2 = h_2(X_{k_1+1},\dots,X_{k_2}),\dots,Y_m = h_m(X_{k_{m-1}},\dots,X_{k_m})$. Then, $Y_1,\dots,Y_m$ are also independent.

Remark 30. In a previous section, we had defined independence of events and now we have defined independence of random variables. How are they related? We leave it to you to check that events $A_1, \ldots, A_n$ are independent (according to the definition in the previous section) if and only if the random variables $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_m}$ are independent (according to the definition in this section).

1. Conditioning of random variables

Let $X_1,\ldots,X_{k+\ell}$ be random variables on a common probability space. Let $f(t_1,\ldots,t_{k+\ell})$ be the pmf of $(X_1,\ldots,X_{k+\ell})$ and let $g(t_1,\ldots,t_\ell)$ be the pmf of $(X_{k+1},\ldots,X_{k+\ell})$ (of course we can compute g from f by summing over the first k indices). Then, for any $s_1,\ldots,s_\ell$ such that $\mathbf{P}\{X_{k+1}=s_1,\ldots,X_{k+\ell}=s_\ell\}>0$, we can define

(1)
$$h_{s_1,\ldots,s_\ell}(t_1,\ldots,t_k)=\mathbf{P}\{X_1=t_1,\ldots,X_k=t_k \,\Big|\, X_{k+1}=s_1,\ldots,X_{k+\ell}=s_\ell\}=\frac{f(t_1,\ldots,t_k,s_1,\ldots,s_\ell)}{g(s_1,\ldots,s_\ell)}.$$

It is easy to see that $h_{s_1,\ldots,s_\ell}(\cdot)$ is a pmf on $\mathbb{R}^k$. It is called the conditional pmf of $(X_1,\ldots,X_k)$ given that $X_{k+1}=s_1,\ldots,X_{k+\ell}=s_\ell$.

Its interpretation is as follows. Originally, we had random observables $X_1,\ldots,X_k$ which had a certain joint pmf. Then, we observe the values of the random variables $X_{k+1},\ldots,X_{k+\ell}$, say they turn out to be $s_1,\ldots,s_\ell$, respectively. Then, we update the distribution (or pmf) of $X_1,\ldots,X_k$ according to the above recipe. The conditional pmf is the new function $h_{s_1,\ldots,s_\ell}$.

Exercise 1. Let $(X_1,\ldots,X_{n-1})$ be a random vector with multinomial distribution with parameters $r,n,p_1,\ldots,p_n$. Let $k<n-1$. Given that $X_{k+1}=s_1,\ldots,X_{n-1}=s_{n-k+1}$, show that the conditional distribution of $(X_1,\ldots,X_k)$ is multinomial with parameters $r',n',q_1,\ldots,q_{k+1}$, where $r'=r-(s_1+\cdots+s_{n-k+1})$, $n'=k+1$, $q_j=p_j/(p_1+\cdots+p_k+p_n)$ for $j\le k$ and $q_{k+1}=p_n/(p_1+\cdots+p_k+p_n)$.

This looks complicated, but is utterly obvious if you think in terms of assigning r balls into n urns by putting each ball into the urns with probabilities $p_1,\ldots,p_n$ and letting X_j denote the number of balls that end up in the j^{th} urn.

Conditional densities: Now, suppose that $X_1,\ldots,X_{k+\ell}$ have joint density $f(t_1,\ldots,t_{k+\ell})$ and let $g(s_1,\ldots,s_\ell)$ be the density of $(X_{k+1},\ldots,X_{k+\ell})$. Then, we define the conditional density of $(X_1,\ldots,X_k)$ given $X_{k+1}=s_1,\ldots,X_{k+\ell}=s_\ell$ as

$$h_{s_1,\ldots,s_\ell}(t_1,\ldots,t_k)=\frac{f(t_1,\ldots,t_k,s_1,\ldots,s_\ell)}{g(s_1,\ldots,s_\ell)}. \tag{2}$$

This is well-defined whenever $g(s_1,\ldots,s_\ell)>0$.

Remark 2. Note the difference between (1) and (2). In the latter, we have left out the middle term because $\mathbf{P}\{X_{k+1}=s_1,\ldots,X_{k+\ell}=s_\ell\}=0$. In (1), the definition of pmf comes from the definition of conditional probability of events, but in (2) this is not so. We simply define the conditional density by analogy with the case of conditional pmf. This is similar to the difference between interpretation of pmf ($f(t)$ is actually the probability of an event) and pdf ($f(t)$ is not the probability of an event, but the density of probability near t).

Example 3. Let (X, Y) have bivariate normal density $f(x, y) = \frac{\sqrt{ab-c^2}}{2\pi} e^{-\frac{1}{2}(ax^2+by^2+2cxy)}$ (so we assume $a > 0, b > 0, ab - c^2 > 0$). We can show that the marginal distribution of Y is $N(0, \frac{a}{ab-c^2})$, i.e., it has density $g(y) = \frac{\sqrt{ab-c^2}}{\sqrt{2\pi a}} e^{-\frac{ab-c^2}{2a}y^2}$. Hence, the conditional density of X given $Y = y$ is

$$h_y(x) = \frac{f(x, y)}{g(y)} = \frac{\sqrt{a}}{\sqrt{2\pi}} e^{-\frac{a}{2}(x+\frac{c}{a}y)^2}.$$

Thus, the conditional distribution of X given $Y = y$ is $N(-\frac{cy}{a}, \frac{1}{a})$. Compare this with marginal (unconditional) distribution of X which is $N(0, \frac{b}{ab-c^2})$.

In the special case when $c = 0$, we see that for any value of y, the conditional distribution of X given $Y = y$ is the same as the unconditional distribution of X. What does this mean? It is just another way of saying that X and Y are independent! Indeed, when $c = 0$, the joint density $f(x, y)$ splits into a product of two functions, one of x alone and one of y alone.

Exercise 4. Let (X, Y) have joint density $f(x, y)$. Let the marginal densities of X and Y be $g(x)$ and $h(y)$, respectively. Let $h_x(y)$ be the conditional density of Y given $X = x$.

(1) If X and Y are independent, show that for any x, we have $h_x(y) = h(y)$ for all y.

(2) If $h_x(y) = h(y)$ for all y and for all x, show that X and Y are independent.

Analogous statements hold for the case of pmf as well.

1. Moments and Moment Generating Function

Moments: Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a p-dimensional random vector of either discrete type, or (absolutely) continuous type. Let $f_{\mathbf{X}}$ and $S_{\mathbf{X}} = \{\mathbf{X} \in \mathbb{R}^p : f_{\mathbf{X}}(\mathbf{x}) > 0\}$ denote the pmf (or, pdf) and support of $\mathbf{X}$ (or, $f_{\mathbf{X}}$). Further, let f_{X_i} and $S_{X_i} = \{x \in \mathbb{R} : f_{X_i}(x) > 0\}$ denote the pmf (or, pdf) and support of X_i (or, f_{X_i}) for $i = 1, \ldots, p$.

Let $\psi : \mathbb{R}^p \to \mathbb{R}$ be a function such that $\mathbf{E}[\psi(\mathbf{X})]$ exists (i.e., $\mathbf{E}|\psi(\mathbf{X})| < \infty$).

- If $\mathbf{X}$ is of discrete type, then

$$\mathbf{E}(\psi(\mathbf{X})) = \sum_{\mathbf{x} \in S_{\mathbf{X}}} \psi(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}).$$

- If $\mathbf{X}$ is of absolutely continuous type, then

$$\mathbf{E}(\psi(\mathbf{X})) = \int_{\mathbb{R}^p} \psi(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

- For non-negative integers $k_1, \ldots, k_p$, let $\psi(\mathbf{x}) = x_1^{k_1} \cdots x_p^{k_p}$. Then,

$$\mu'_{k_1,\ldots,k_p} = \mathbf{E}\left(X_1^{k_1} \cdots X_p^{k_p}\right)$$

is called a joint raw moment of order $k_1 + \cdots + k_p$ of $\mathbf{X}$.

- For non-negative integers $k_1, \ldots, k_p$, let $\psi(\mathbf{x}) = (x_1 - \mathbf{E}(X_1))^{k_1} \cdots (x_p - \mathbf{E}(X_p))^{k_p}$. Then

$$\mu_{k_1,\ldots,k_p} = \mathbf{E}\left((X_1 - \mathbf{E}(X_1))^{k_1} \cdots (X_p - \mathbf{E}(X_p))^{k_p}\right)$$

is called a joint central moment of order $k_1 + \cdots + k_p$ of $\mathbf{X}$.

- Let $\psi(\mathbf{x}) = (x_i - \mathbf{E}(X_i))(x_j - \mathbf{E}(X_j))$ for $i, j = 1, \ldots, p$. Then, the covariance between X_i and X_j is

$$\begin{aligned}\operatorname{Cov}(X_i, X_j) &= \mathbf{E}\left((X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))\right) \\ &= \mathbf{E}(X_i X_j) - \mathbf{E}(X_i)\mathbf{E}(X_j).\end{aligned}$$

Let $\mathbf{X} = (X_1, X_2, \ldots, X_{p_1})$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{p_2})$ be random vectors, and let $a_1, \ldots, a_{p_1}$ and $b_1, \ldots, b_{p_2}$ be real constants. Assume that the involved expectations exist. Then,

(i) $\mathbf{E}\left(\sum_{i=1}^{p_1} a_i X_i\right) = \sum_{i=1}^{p_1} a_i \mathbf{E}(X_i)$

(ii) $\operatorname{Cov}\left(\sum_{i=1}^{p_1} a_i X_i, \sum_{j=1}^{p_2} b_j Y_j\right) = \sum_{i=1}^{p_1}\sum_{j=1}^{p_2} a_i b_j \operatorname{Cov}(X_i, Y_j)$.

In particular,

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{p_1} a_i X_i\right) &= \sum_{i=1}^{p_1} a_i^2 \operatorname{Var}(X_i) + \sum_{i=1}^{p_1} \sum_{\substack{j=1 \\ j\neq i}}^{p_1} a_i a_j \operatorname{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{p_1} a_i^2 \operatorname{Var}(X_i) + 2 \sum_{1\leq i<j\leq p_1} a_i a_j \operatorname{Cov}(X_i, X_j).\end{aligned}$$

We now state a property of expectations related to independence.

Lemma 1. *Let X, Y be random variables on a common probability space. If X and Y are independent, then $\mathbf{E}[H_1(X)H_2(Y)] = \mathbf{E}[H_1(X)]\mathbf{E}[H_2(Y)]$ for any functions $H_1, H_2 : \mathbb{R} \to \mathbb{R}$ (for which the expectations exist). In particular, $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.*

Proof. Independence means that the joint density (analogous statements for pmf omitted) of (X, Y) is of the form $f(t, s) = g(t)h(s)$, where $g(t)$ is the density of X and $h(s)$ is the density of Y. Hence,

$$\mathbf{E}[H_1(X)H_2(Y)] = \iint H_1(t)H_2(s)f(t,s)dtds = \left(\int_{-\infty}^{\infty} H_1(t)g(t)dt\right)\left(\int_{-\infty}^{\infty} H_2(s)h(s)ds\right)$$

which is precisely $\mathbf{E}[H_1(X)]\mathbf{E}[H_2(Y)]$. ■

Moment Generating Function: Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a p-dimensional random vector, and

$$A = \left\{\mathbf{t} = (t_1, t_2, \ldots, t_p) \in \mathbb{R}^p : \mathbf{E}\left(e^{\sum_{i=1}^{p} t_i X_i}\right) < \infty\right\}.$$

Define the function $M_{\mathbf{X}} : A \to \mathbb{R}$ by

$$M_{\mathbf{X}}(\mathbf{t}) = \mathbf{E}\left(e^{\sum_{i=1}^{p} t_i X_i}\right), \quad \mathbf{t} = (t_1, t_2, \ldots, t_p) \in A.$$

The function $M_{\mathbf{X}} : A \to \mathbb{R}$ is called the joint moment generating function (mgf) of random vector $\mathbf{X}$. For $\mathbf{a} = (a_1, a_2, \ldots, a_p) \in \mathbb{R}^p$, $-\mathbf{a} = (-a_1, -a_2, \ldots, -a_p)$ and $(-\mathbf{a}, \mathbf{a}) = \{\mathbf{t} \in \mathbb{R}^p : -a_i < t_i < a_i \text{ for } i = 1, \ldots, p\}$. As in the one-dimensional case, many properties of probability distribution of $\mathbf{X}$ can be studied through the joint mgf of $\mathbf{X}$. Some of the results, which may be useful in this direction, are provided below (without their proofs). Note that $M_{\mathbf{X}}(0_p) = 1$, where 0_p is the vector of 0s.

If $X_1, \ldots, X_p$ are independent, then

$$M_{\mathbf{X}}(\mathbf{t}) = \mathbf{E}\left(e^{\sum_{i=1}^{p} t_i X_i}\right) = \mathbf{E}\left(\prod_{i=1}^{p} e^{t_i X_i}\right) = \prod_{i=1}^{p} \mathbf{E}\left(e^{t_i X_i}\right) = \prod_{i=1}^{p} M_{X_i}(t_i) \text{ for } \mathbf{t} \in \mathbb{R}^p.$$

Suppose that $M_{\mathbf{X}}(\mathbf{t})$ exists in a rectangle $(-\mathbf{a}, \mathbf{a}) \subseteq \mathbb{R}^p$. Then, $M_{\mathbf{X}}(\mathbf{t})$ possesses partial derivatives of all orders in $(-\mathbf{a}, \mathbf{a})$. Furthermore, for positive integers $k_1, \ldots, k_p$

$$\mathbf{E}\left(X_1^{k_1} X_2^{k_2} \cdots X_p^{k_p}\right) = \left[\frac{\partial^{k_1+k_2+k_3+\cdots+k_p}}{\partial t_1^{k_1} \cdots \partial t_p^{k_p}} M_{\mathbf{X}}(\mathbf{t})\right]_{\mathbf{t}=0_p}.$$

For $i \neq j$ with $i, j \in \{1, \ldots, p\}$, define

$$\begin{aligned}\operatorname{Cov}(X_i, X_j) &= \mathbf{E}(X_i X_j) - \mathbf{E}(X_i)\,\mathbf{E}(X_j) \\ &= \left[\frac{\partial^2}{\partial t_i \partial t_j} M_{\mathbf{X}}(\mathbf{t})\right]_{\mathbf{t}=0_p} - \left[\frac{\partial}{\partial t_i} M_{\mathbf{X}}(\mathbf{t})\right]_{\mathbf{t}=0_p} \left[\frac{\partial}{\partial t_j} M_{\mathbf{X}}(\mathbf{t})\right]_{\mathbf{t}=0_p} \\ &= \left[\frac{\partial^2}{\partial t_i \partial t_j} \Psi_{\mathbf{X}}(\mathbf{t})\right]_{\mathbf{t}=0_p},\end{aligned}$$

where $\Psi_{\mathbf{X}}(\mathbf{t}) = \ln M_{\mathbf{X}}(\mathbf{t})$.
We also have $M_{\mathbf{X}}(0, \ldots, 0, t_i, 0, \ldots, 0, t_j, 0, \ldots, 0) = \mathbf{E}\left(e^{t_i X_i + t_j X_j}\right) = M_{X_i, X_j}(t_i, t_j)$ for $i, j \in \{1, \ldots, p\}$.

Identically distributed: Let $\mathbf{X}$ and $\mathbf{Y}$ be two p-dimensional random vectors, defined on the same probability space. Then, $\mathbf{X}$ and $\mathbf{Y}$ are said to have the same distribution (written as $\mathbf{X} \overset{D}{=} \mathbf{Y}$) if

$$F_{\mathbf{X}}(\mathbf{x}) = F_{\mathbf{Y}}(\mathbf{x})\ \forall\ \mathbf{x} \in \mathbb{R}^p \text{ (i.e., they have the same distribution function).}$$

If $\mathbf{X}$ and $\mathbf{Y}$ are p-dimensional random vectors of discrete type with joint pmf $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$, respectively. Then, $\mathbf{X} \overset{D}{=} \mathbf{Y}$ if and only if $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(\mathbf{x})\ \forall\ \mathbf{x} \in \mathbb{R}^p$.

If $\mathbf{X}$ and $\mathbf{Y}$ are p-dimensional random vectors of absolutely continuous type with joint pdf $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$, respectively. Then, $\mathbf{X} \overset{D}{=} \mathbf{Y}$ if and only if $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(\mathbf{x})\ \forall\ \mathbf{x} \in \mathbb{R}^p$.

Let $\mathbf{X}$ and $\mathbf{Y}$ be p-dimensional random vectors with $\mathbf{X} \overset{D}{=} \mathbf{Y}$. Then, for any function $h : \mathbb{R}^p \to \mathbb{R}$, we have

(i) $h(\mathbf{X}) \overset{D}{=} h(\mathbf{Y})$,

(ii) $\mathbf{E}[h(\mathbf{X})] = \mathbf{E}[h(\mathbf{Y})]$ (provided the expectations exist).

Let $\mathbf{X}$ and $\mathbf{Y}$ be two random vectors having mgfs $M_{\mathbf{X}}$ and $M_{\mathbf{Y}}$ that are finite on a rectangle $(-\mathbf{a}, \mathbf{a})$ for some $\mathbf{a} = (a_1, a_2, \ldots, a_p) \in \mathbb{R}^p$. Suppose that

$$M_{\mathbf{X}}(\mathbf{t}) = M_{\mathbf{Y}}(\mathbf{t}) \quad \forall \quad \mathbf{t} \in (-\mathbf{a}, \mathbf{a}).$$

Then, $\mathbf{X} \overset{D}{=} \mathbf{Y}$.

If $X_1, X_2, \ldots, X_p$ are independent and identically distributed (i.i.d.), i.e., $X_i \overset{D}{=} X_1$ for $i = 2, \ldots, p$, then

$$M_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^{p} M_{X_1}(t_i) \text{ for } \mathbf{t} \in \mathbb{R}^p.$$

Define $Y = \sum_{i=1}^{p} X_i$ and $\bar{X} = \frac{1}{p} \sum_{i=1}^{p} X_i$, then

$$M_Y(t) = [M_{X_1}(t)]^p \text{ and } M_{\bar{X}}(t) = [M_{X_1}(t/p)]^p \text{ for } t \in \mathbb{R}.$$

Exercise 2. Let $X_1, X_2, \ldots, X_p$ be independent random variables such that $X_i \sim N(\mu_i, \sigma_i^2)$ with $-\infty < \mu_i < \infty$ and $\sigma_i > 0$ for $i = 1, \ldots, p$. If $a_1, \ldots, a_p$ are real constants (such that not all of them are zero), then show that

$$\sum_{i=1}^{p} a_i X_i \sim N\left(\sum_{i=1}^{p} a_i \mu_i, \sum_{i=1}^{p} a_i^2 \sigma_i^2\right).$$

2. Covariance and Correlation

Covariance: Let X, Y be random variables on a common probability space. The *covariance* of X and Y is defined as $\text{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$. It can also be written as $\text{Cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$.

Correlation: Let X, Y be random variables on a common probability space. Their *correlation* is defined as $\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$.

Measures of association: The marginal distributions of X and Y do not determine the joint distribution of (X,Y). In particular, giving the means and standard deviations of X and Y does not tell anything about possible relationships between the two.
Read this: http://probability.ca/jeff/teaching/uncornor.html.

Covariance is the quantity that is used to measure the "association" of Y and X. Correlation is a dimension free quantity that measures the same. For example, we shall see that if $Y = X$, then $\text{Corr}(X,Y) = +1$, if $Y = -X$ then $\text{Corr}(X,Y) = -1$. Further, if X and Y are independent, then $\text{Corr}(X,Y) = 0$. In general, if an increase in X is likely to mean an increase in Y, then the correlation is positive and if an increase in X is likely to mean a decrease in Y then the correlation is negative.

Properties of covariance and variance: Let X, Y, X_i, Y_i be random variables on a common probability space. Small letters a, b, c etc. will denote scalars.

(1) (Bilinearity): $\text{Cov}(aX_1 + bX_2, Y) = a\text{Cov}(X_1, Y) + b\text{Cov}(X_2, Y)$ and $\text{Cov}(X, aY_1 + bY_2) = a\text{Cov}(X, Y_1) + b\text{Cov}(X, Y_2)$.

(2) (Symmetry): $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.

(3) (Positivity): $\text{Cov}(X,X) \geq 0$ with equality if and only if X is a constant random variable. Indeed, $\text{Cov}(X,X) = \text{Var}(X)$.

Exercise 3. If X and Y are independent, then show that $\text{Cov}(X,Y) = 0$ and hence, $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.

However, $\text{Cov}(X,Y) = 0$ does not necessarily imply that X and Y are independent!

Example 4. Let $\mathbf{X} = (X_1, X_2)$ be a bivariate random vector of absolutely continuous type with pdf given by

$$f_{\mathbf{X}}(x_1, x_2) = \begin{cases} 1, & \text{if } 0 < |x_2| \leq x_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Then,

$$\mathbf{E}\left(X_1X_2\right) = \int_0^1 \int_{-x_1}^{x_1} x_1 x_2 \,\mathrm{d}x_2\,\mathrm{d}x_1 = 0,$$

$$\mathbf{E}\left(X_1\right) = \int_0^1 \int_{-x_1}^{x_1} x_1 \,\mathrm{d}x_2\,\mathrm{d}x_1 = \frac{2}{3},$$

$$\mathbf{E}\left(X_2\right) = \int_0^1 \int_{-x_1}^{x_1} x_2 \,\mathrm{d}x_2\,\mathrm{d}x_1 = 0,$$

and

$$\mathrm{Cov}\left(X_1, X_2\right) = \mathbf{E}\left(X_1X_2\right) - \mathbf{E}\left(X_1\right)\mathbf{E}\left(X_2\right) = 0$$

Therefore,

$$\mathrm{Corr}\left(X_1, X_2\right) = 0$$

i.e., X_1 and X_2 are uncorrelated.

Exercise 5. Show that

$$f_{\mathbf{X}}\left(x_1, x_2\right) \neq f_{X_1}\left(x_1\right) f_{X_2}\left(x_2\right) \quad \forall\ \left(x_1, x_2\right) \in \mathbb{R}^2.$$

Therefore, X_1 and X_2 are not independent.

Remark 6. Note that the properties of covariance are very much like properties of inner-products in vector spaces. In particular, we have the following analogue of the well-known inequality for vectors $(\mathbf{u}\cdot\mathbf{v})^2 \leq (\mathbf{u}\cdot\mathbf{u})(\mathbf{v}\cdot\mathbf{v})$.

Cauchy-Schwarz inequality: If X and Y are random variables with finite variances, then

$$(\mathrm{Cov}(X,Y))^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y)$$

with equality if and only if $Y = aX + b$ for some scalars a, b.
Follow the proof of Cauchy-Schwarz inequality that you have seen for vectors. This just means that $\mathrm{Var}(X + tY) \geq 0$ for any scalar t and choose an appropriate t to get the Cauchy-Schwarz inequality.

3. Multivariate Normal Distribution

Bivariate Normal Distribution: We say that (X, Y) follows a bivariate normal distribution if its pdf is given by

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-Q/2} \text{ for } \quad -\infty < x < \infty, \quad -\infty < y < \infty,$$

where

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]$$

with $-\infty < \mu_i < \infty, \sigma_i > 0$ for $i = 1, 2$, and ρ satisfies $\rho^2 < 1$. Clearly, this function is positive everywhere in $\mathbb{R}^2$.

Remark 7. Note that we can derive this new formulation from the earlier one by defining

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Exercise 8. Compute Σ^{-1} and $det(\Sigma)$.

This pdf has mgf given by (proof is given later):

$$M_{(X,Y)}(t_1, t_2) = \exp\left\{t_1\mu_1 + t_2\mu_2 + \frac{1}{2}\left(t_1^2\sigma_1^2 + 2t_1t_2\rho\sigma_1\sigma_2 + t_2^2\sigma_2^2\right)\right\}.$$

Thus, the mgf of X is

$$M_X(t_1) = M_{(X,Y)}(t_1, 0) = \exp\left\{t_1\mu_1 + \frac{1}{2}t_1^2\sigma_1^2\right\},$$

while the mgf of Y is

$$M_Y(t_2) = M_{(X,Y)}(0, t_2) = \exp\left\{t_2\mu_2 + \frac{1}{2}t_2^2\sigma_2^2\right\}.$$

Hence, X has a $N\left(\mu_1, \sigma_1^2\right)$ distribution. In the same way, Y has a $N\left(\mu_2, \sigma_2^2\right)$ distribution. Thus, μ_1 and μ_2 are the respective means of X and Y, while σ_1^2 and σ_2^2 are the respective variances of X and Y.

Exercise 9. For the parameter ρ, show that

$$\mathbf{E}(XY) = \frac{\partial^2 M_{(X,Y)}}{\partial t_1 \partial t_2}(0,0) = \rho\sigma_1\sigma_2 + \mu_1\mu_2.$$

Hence, $\mathrm{Cov}(X,Y) = \rho\sigma_1\sigma_2$. As the notation suggests, ρ is the correlation coefficient between X and Y (Check!).

Remark 10. We know that if X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$. For the bivariate normal distribution, if $\rho = 0$ (equivalently, $\mathrm{Cov}(X,Y) = 0$), then the joint mgf of (X,Y) factors into the product of the marginal mgfs. Hence, X and Y are independent random variables. Thus, if (X,Y) has a bivariate normal distribution, then X and Y are independent if and only if they are uncorrelated (i.e., $\rho = 0$). Read below:
https://en.wikipedia.org/wiki/Normally_distributed_and_uncorrelated_does_not_imply_independent.

Check the file 'N_2distribution.R'.

Multivariate Normal Distribution: In this section, we generalize the bivariate normal distribution to the n-dimensional multivariate normal distribution. The derivation of the distribution is simplified by first discussing the standardized variable case, and then proceeding to the general case. Consider the random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$, where $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$ random variables. Then, the density of $\mathbf{Z}$ is

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z_i^2\right\} = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} z_i^2\right\} = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{-\frac{1}{2}\mathbf{z}'\mathbf{z}\right\}$$

for $\mathbf{z} \in \mathbb{R}^n$.

Definition 11. We define $\mathbf{E}[\mathbf{X}]$ as the n-dimensional vector $(\mathbf{E}[X_1], \ldots, \mathbf{E}[X_n])'$, and $\mathrm{Cov}[\mathbf{X}]$ as a $n \times n$ matrix with the (i,j)th element as $\mathrm{Cov}(X_i, X_j)$ for $1 \le i, j \le n$.

Note that the diagonal elements of $\mathrm{Cov}[\mathbf{X}]$ (a symmetric matrix) are the componentwise variances Var$[X_i]$ for $1 \le i \le n$.

Exercise 12. The mean and covariance matrix of $\mathbf{Z}$ are

$$\mathbf{E}[\mathbf{Z}] = 0_n \text{ and } \mathrm{Cov}[\mathbf{Z}] = I_n,$$

where 0_n is the vector of 0s and I_n denotes the identity matrix of order n.

The mgf of Z_is evaluated at t_i is $\exp\left\{t_i^2/2\right\}$ for $i = 1, \ldots, n$. Since the Z_is are independent, the mgf of $\mathbf{Z}$ is

$$M_{\mathbf{Z}}(\mathbf{t}) = \mathbf{E}\left[\exp\left\{\mathbf{t}'\mathbf{Z}\right\}\right] = \mathbf{E}\left[\prod_{i=1}^{n} \exp\left\{t_i Z_i\right\}\right] = \prod_{i=1}^{n} \mathbf{E}\left[\exp\left\{t_i Z_i\right\}\right] = \exp\left\{\frac{1}{2}\sum_{i=1}^{n} t_i^2\right\} = \exp\left\{\frac{1}{2}\mathbf{t}'\mathbf{t}\right\}.$$

for all $\mathbf{t} \in \mathbb{R}^n$. We say that $\mathbf{Z}$ has a multivariate normal distribution with mean vector 0_n and covariance matrix I_n. We abbreviate this by saying that $\mathbf{Z}$ has a $N_n\left(0_n, I_n\right)$ distribution.

For the *general case*, suppose Σ is a $n \times n$ symmetric and positive definite matrix. Then, from linear algebra, we can always decompose Σ as follows:

$$\Sigma = \Gamma' \Lambda \Gamma,$$

where Λ is the diagonal matrix $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ satisfying $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ are the eigenvalues and the columns of Γ' (say, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$) are the corresponding eigenvectors of Σ. This decomposition is called the spectral decomposition of Σ. The matrix Γ is orthogonal, i.e., $\Gamma^{-1} = \Gamma'$ and hence, $\Gamma\Gamma' = I_n$.

We define the square root of the positive definite matrix Σ as

$$\Sigma^{1/2} = \Gamma' \Lambda^{1/2} \Gamma.$$

Note that $\Sigma^{1/2}$ is symmetric and positive definite. It is now easy to show that

$$\left(\Sigma^{1/2}\right)^{-1} = \Gamma' \Lambda^{-1/2} \Gamma.$$

We write the left side of this equation as $\Sigma^{-1/2}$. Suppose $\mathbf{Z}$ has a $N_n(0_n, I_n)$ distribution. Let Σ be a positive definite, symmetric matrix and μ be an $n \times 1$ vector of constants. Define the random vector $\mathbf{X}$ by

$$\mathbf{X} = \Sigma^{1/2}\mathbf{Z} + \mu.$$

We now have

$$\mathbf{E}[\mathbf{X}] = \mu \text{ and } \operatorname{Cov}[\mathbf{X}] = \Sigma^{1/2}\Sigma^{1/2} = \Sigma.$$

Further, the mgf of $\mathbf{X}$ is given by

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) = \mathbf{E}\left[\exp\left\{\mathbf{t}'\mathbf{X}\right\}\right] &= \mathbf{E}\left[\exp\left\{\mathbf{t}'\Sigma^{1/2}\mathbf{Z} + \mathbf{t}'\mu\right\}\right] \\ &= \exp\left\{\mathbf{t}'\mu\right) \left[\exp\left\{\left(\Sigma^{1/2}\mathbf{t}\right)'\mathbf{Z}\right\}\right] \\ &= \exp\left\{\mathbf{t}'\mu\right\} \exp\left\{(1/2)\left(\Sigma^{1/2}\mathbf{t}\right)'\Sigma^{1/2}\mathbf{t}\right\} \\ &= \exp\left\{\mathbf{t}'\mu\right\} \exp\left\{(1/2)\mathbf{t}'\Sigma\mathbf{t}\right\} \\ &= \exp\left\{\mathbf{t}'\mu + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}\right\}. \end{aligned}$$

The transformation between $\mathbf{X}$ and $\mathbf{Z}$ is one-to-one with the inverse transformation

$$\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \mu)$$

and Jacobian $\left|\Sigma^{-1/2}\right| = |\Sigma|^{-1/2}$. Hence, upon simplification, the pdf of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\right\} \quad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

The following theorem says that a linear transformation of a multivariate normal random vector has a multivariate normal distribution.

Theorem 13. *Suppose* $\mathbf{X}$ *has a* $N_n(\mu, \Sigma)$ *distribution. Let* $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$*, where* A *is an* $m \times n$ *matrix and* $\mathbf{b} \in \mathbb{R}^m$*. Then,* $\mathbf{Y}$ *has a* $N_m(A\mu + \mathbf{b}, A\Sigma A')$ *distribution.*

Proof. For $\mathbf{t} \in \mathbb{R}^m$, the mgf of $\mathbf{Y}$ is

$$\begin{aligned}
M_{\mathbf{Y}}(\mathbf{t}) &= \mathbf{E}\left[\exp\left\{\mathbf{t}'\mathbf{Y}\right\}\right] \\
&= \mathbf{E}\left[\exp\left\{\mathbf{t}'(A\mathbf{X} + \mathbf{b})\right\}\right] \\
&= \exp\left\{\mathbf{t}'\mathbf{b}\right\} \mathbf{E}\left[\exp\left\{\left(A'\mathbf{t}\right)' \mathbf{X}\right\}\right] \\
&= \exp\left\{\mathbf{t}'\mathbf{b}\right\} \exp\left\{\left(A'\mathbf{t}\right)' \mu + (1/2)\left(A'\mathbf{t}\right)' \Sigma \left(A'\mathbf{t}\right)\right\} \\
&= \exp\left\{\mathbf{t}'(A\mu + \mathbf{b}) + \frac{1}{2}\mathbf{t}' A\Sigma A'\mathbf{t}\right\}
\end{aligned}$$

which is the mgf of a $N_m(A\mu + \mathbf{b}, A\Sigma A')$ distribution. ■