Jimmy Yang

MATH 318 Notes

Notes for MATH 318: Probability with Physical Applications taught by Dr. Omer Angel.
Contains additional stuff and examples from other places that are not covered during classes.
If you have any questions or comments related to the notes, free to email me at
marethyu@student.ubc.ca.
Last modified: February 25, 2023

Contents

1   Counting

1.1   Permutations

Permutations are combinations where order matters. How permutations are calculated depends on whether repetitions are allowed or not. The number of permutations made from choosing k from n items when repetitions are allowed is n^k.

> EXAMPLE 1.
>
> Find the number of permutations of ab when letters can be repeated.
>
> The permutations are: aa, ab, ba, bb. There is a total of $2^2 = 4$ permutations.

When repetitions are not allowed, the number of permutations is just $n!$. Here is one way to understand where this formula came from: for the first item, there are n choices to place and since no repetitions are allowed, there are $n-1$ choices left for the second item and so on.

> EXAMPLE 2.
>
> An Olympic final consists of two runners from each of four different countries. An ordered selection of three of the eight runners win the medals (gold, silver, and bronze). How many different possibilities are there if the medal winners are all from different countries?
>
> We are looking for an ordered selection of winners. There are 8 choices for the gold medal winner. After the winner is chosen, since we want winners from different countries, there are 6 options left for the solver medal winner. This leaves with 4 options for the bronze medal winner. The total number of possible selections is therefore $8 \times 6 \times 4 = 192$.

> EXAMPLE 3.
>
> Find the number of permutations of abc.
>
> The permutations are: $abc, acb, bac, bca, cab, cba$. There are total of $3(2!) = 3! = 6$ permutations. Three unique letters times the number of permutations of remaining letters. If there are 4 letters, the answer would be $4(3!) = 24$.
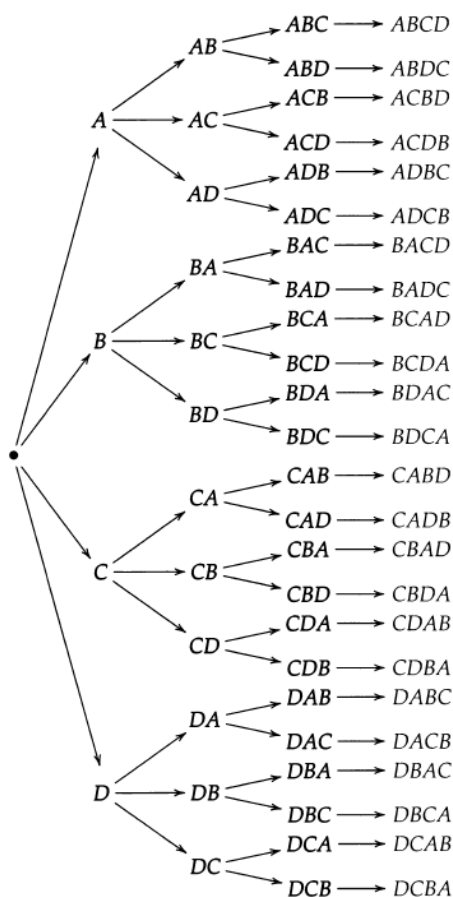
Figure 1: Permutations of $ABCD$ visualized

The above figure shows how to generate permutations of $ABCD$. We start with 4 choices for the first letter. Once we got the first letter, there are three choices remaining for the second letter and so on.

EXAMPLE 4.

Find the number of words created from rearranging letters $ALLELE$.

There are two Es and three Ls. Notice that there is no difference when you reorder these letters. Each word created from two Es and three Ls corresponds to 2!3! permutations.

The number of permutations is 6! which includes duplicates. To get the final answer, we divide it by 2!3! which gives $\frac{6!}{2!3!} = 60$ total words.

EXAMPLE 5.

Calculate the probability of obtaining exactly 3 heads and tails for a total of 6 coin tosses.

There are the total of $2^6 = 64$ possible outcomes. The number of times when the combination $HHHTTT$ occurs is $\frac{6!}{3!3!} = 20$. Therefore the probability is $\frac{20}{64} = \frac{5}{16}$.

To calculate the number of permutations made from choosing k out of n items with no repetition:

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

It is usually denoted by $_nP_k$.

1.2 Combinations

EXAMPLE 6.

How many ways to pick two out of $\{a, b, c, d\}$?

Initially, there are four choices for the first letter of an ordered pair. After picking one, there are three choices left. There are $4(3) = 12$ ordered pairs: $ab, ba, ac, ca, ad, da, bc, cb, bd, db, cd, dc$. However, we don't care about the order (ie. ab and ba is considered the same) so we divide by 2! which gives: $\frac{12}{2} = 6$ combinations.

To pick k out of n items, the total number of combinations is the number of ways to create ordered combinations (permutations) of size k divided by the number of permutations of k items in combinations. The calculation process is outlined below:

$$\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \frac{\frac{n!}{(n-k)!}}{k!} = \frac{n!}{k!(n-k)!}$$

We use $\binom{n}{k}$ to denote $\frac{n!}{k!(n-k)!}$. It is read as "n choose k." This is a binomial coefficient. Another way to think about combinations is that it is the same as finding how many ways to rearrange n cards where k cards are coloured red and $n-k$ cards are coloured green. Also note that:

$$\binom{n}{n-k} = \frac{n!}{(n-k)!(n-(n-k))!} = \frac{n!}{(n-k)!k!} = \binom{n}{k}$$

$\binom{n}{k}$ is maximized when $k = n/2$.

EXAMPLE 7.

How many ways to deal 52 cards to 4 players?

We count the number of ways to pick 13 cards from the deck of 52 cards. We then multiply it by the number of ways to pick another set of 13 cards from the deck of $52 - 13 = 39$ cards. We multiply it again by the number of ways to pick another set of 13 cards from the deck of $39 - 13 = 26$ cards. The remaining number of cards is $26 - 13 = 13$ which will be given to the last player.

$$\binom{52}{13}\binom{39}{13}\binom{26}{13}\binom{13}{13} = \frac{52!}{(13!)^4}$$

There is an interesting observation that can be used to calculate binomial coefficients. Suppose you toss a coin n times and k of n tosses are heads. The number of possible ways to get k heads out of n tosses is given by $\binom{n}{k}$. To find this binomial coefficient, consider two cases: the first toss is H and the first toss is not H. When the first toss is H, the remaining tosses $n-1$ must have $k-1$ heads and the number of possible ways it can happen is $\binom{n-1}{k-1}$. When the first toss is not H, the remaining tosses $n-1$ must have k heads and the number of possible ways it can happen is $\binom{n-1}{k}$. The number of possible ways to get k heads out of n tosses is the number of possible ways to get H as the first toss plus the number of possible ways not to get H as the first toss:

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

This is called "Pascal's rule." It can be used to construct Pascal's triangle.

$$1$$

$$1 \qquad 1$$

$$1 \qquad 2 \qquad 1$$

$$1 \qquad 3 \qquad 3 \qquad 1$$

$$1 \qquad 4 \qquad 6 \qquad 4 \qquad 1$$

$\binom{n}{k}$ gives the value at nth row and kth column (both zero-indexed) of the triangle. For example, in the third row and second column, $\binom{2}{1} = 2$. Notice that the total in each row is always 2^n.

1.3 Multinomial Coefficient

The binomial coefficient formula $\binom{n}{k}$ can also be thought of as the number of ways to distribute n items into two boxes in such way that one box has k items and the another has $n - k$ items. Multinomial coefficient generalizes this fact to multiple boxes.

The formula for the multinomial coefficient can be derived from binomial coefficients. Suppose you have r boxes where you want to distribute (distinct) n items. You want each box to have $k_1, k_2, k_3, \ldots, k_r$ items with $k_1 + k_2 + \cdots + k_r = n$. You first count the number of ways to distribute k_1 items in the first box with $n - k_1$ items remaining and this gives $\binom{n}{k_1}$. Then using the remaining $n - k_1$ items to count the number of ways to distribute k_2 items in the second box which gives $\binom{n-k_2}{k_2}$. There are $n - k_1 - k_2$ items left and this will be used to calculate the number of ways to distribute k_3 items into the third box and so on. The number of ways to distribute n items into m boxes is therefore:

$$\binom{n}{k_1}\binom{n - k_1}{k_2}\binom{n - k_1 - k_2}{k_3} \cdots \binom{k_{r-1} + k_r}{k_{r-1}}\binom{k_r}{k_r}$$

Simplifying this expression gives multinomial coefficient:

$$\boxed{\binom{n}{k_1, k_2, k_3, \ldots, k_r} = \frac{n!}{k_1! k_2! k_3! \cdots k_r!}}$$

This can also be interpreted as the number of words if letters appear $k_1, k_2, k_3, \ldots, k_r$ times. The fourth example can also be solved with $\binom{6}{1,2,3}$.

> EXAMPLE 8.
>
> How many ways to arrange the sequence of letters $ABRACADABRA$ in such way that there are no consecutive As?
>
> The position of As must be five numbers from $\{1, 2, 3, \ldots, 11\}$ with no two consecutive numbers. Denote positions of each of the five As by i, j, k, l, m and $i, j-1, k-2, l-3, m-4$ should not contain repetitions. This means $i, j - 1, k - 2, l - 3, m - 4$ should take five numbers from $\{1, 2, 3, \ldots, 7\}$. There are $\binom{7}{5} = 21$ possible sets of locations for the As. For each of them, there are $\binom{6}{4,4} = 180$ arrangements for the remaining letters. The overall answer is $21(180) = 3780$.

1.4   Stars and Bars

A graphical aid involving stars and bars can solve a wide range of problems. Let's walk over the following examples.

EXAMPLE 9.

There are 5 buckets of ice cream, each with different flavour. How many ways to make 3 scoops, provided repetitions are allowed?

It is easier to solve this problem by considering this related problem. We first line up 5 stars like this:

$$\star \quad \star \quad \star \quad \star \quad \star$$

We then replace a star with 3 bars:

$$\star \quad \star \quad \star \quad ||| \quad \star$$

A star means "skip" and a bar means "serve." In this example, the first three ice cream buckets are skipped and all three scoops are from the fourth ice cream bucket. We reduced the original problem is reduced to: "how many ways to arrange 4 stars and 3 bars in $4 + 3 = 7$ places?" This gives $\binom{7}{3} = 35$ ways.

Number of combinations made from choosing k items from n when duplicates are allowed is summarized by the below formula:

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

EXAMPLE 10.

How many different possible face value combinations created from rolling 3 dice?

This is similar to the ice cream problem we did earlier but with 6 different ice cream choices. The answer is $\binom{6+3-1}{3} = 56$.

EXAMPLE 11.

How many solutions does the equation $a + b + c = 12$ have, where a, b, c are non-negative integers?

We first line up 12 stars like this:

$$\star \quad \star \quad \star \quad \star \quad \star \quad \star \quad \star \quad \star \quad \star \quad \star \quad \star \quad \star$$

We then insert 2 bars between the stars, for example:

$$\star \quad \star \quad \star \quad | \star \quad \star \quad \star \quad \star \quad | \star \quad \star \quad \star \quad \star \quad \star$$

Here, it represents one solution: $3 + 4 + 5 = 12$. We reduced the original problem is reduced to: "how many ways to arrange 12 stars and 2 bars in $12 + 2 = 14$ places?" The answer is $\binom{14}{2} = 91$ solutions.

---

2 Probability

2.1 Terminology

- **Sample space** (denoted by S or Ω): A set of all possible outcomes of an experiment. For example, if an experiment involves rolling a die, its sample space is: $S = \{1, 2, 3, 4, 5, 6\}$. Another example: for an experiment involving drawing a set of five cards from a deck of 52 cards, its sample space consists of all possible sets of five cards and its size is $|S| = \binom{52}{5}$.

- **Event** (denoted by E): An event is any subset of sample space: $E \subset S$. The event E occurs when the outcome of the experiment is a member of E. In the dice example, the event would be something like "number less than 4 appear on the roll": $E = \{1, 2, 3\}$. In the poker example, E would be a set of outcomes where there is a full house.

Consider two events E_1 and E_2. Union of events $E_1 \cup E_2$ consists of outcomes that are either in E_1 or E_2 or both. Intersection $E_1 \cap E_2$ consists of outcomes which are both in E_1 and E_2. Events are said to be disjoint (or mutually exclusive) iff $E_1 \cap E_2 = \emptyset$. The complement of the event E is denoted by E^c. It is the set of outcomes in S that is not in E.

Events follow distributive laws: $E_1 \cap (E_2 \cup E_3) = (E_1 \cap E_2) \cup (E_1 \cap E_3)$ and $E_1 \cup (E_2 \cap E_3) = (E_1 \cup E_2) \cap (E_1 \cup E_3)$.

Revisiting the dice example, let E_1 be the event when a number less than 4 appears on the roll: $E_1 = \{1, 2, 3\}$ and let E_2 be the event when an odd number appears on the roll: $E_2 = \{1, 3, 5\}$. Union of the these events is: $E_1 \cup E_2 = \{1, 2, 3, 5\}$ and the intersection is: $E_1 \cap E_2 = \{1, 3\}$. The complement of E_1 is $E_1^c = \{4, 5, 6\}$.
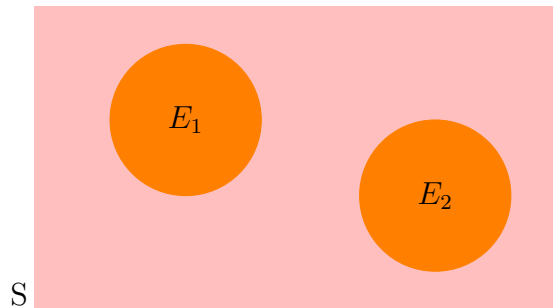

Figure 2: Cartoon illustrating sample space S and disjoint events E_1 and E_2

2.2 Defining Probability

When repeating the experiment with sample space S n times, let $N_n(E)$ be the number of times E occurs. Probability $P(E)$ of the event E is defined as:

$$P(E) = \lim_{n \to \infty} \frac{N_n(E)}{n}$$

More formally, Probability is defined as a measure which is a function P that assigns to each E as a number and satisfies the following axioms:

 i $0 \le P(E) \le 1$ for all E

 ii $P(S) = 1$ and $P(\emptyset) = 0$

iii For disjoint events $E_1, E_2, ..., E_N$: $P\left(\bigcup\limits_{n=1}^{N} E_n\right) = \sum\limits_{n=1}^{N} P(E_n)$

If S is finite (which happens often), we care about finite sequence of events only. The tuple $(S, \mathcal{F}, P)$ constructs a probability space where $\mathcal{F} = \mathcal{P}(S)^1$ is an event space which consists of all possible outcomes in the sample space S. Each event is often equally likely but not always (see the coin-flipping example below for instance). When events are equally likely, $P(E) = \frac{|E|}{|S|}$.

EXAMPLE 12.

A coin is tossed until either two Tails appear successively, or until the fifth toss, whichever comes first. The outcome is the resulting sequence of coins.

There are 20 possible outcomes which is described by the following sample space

$$S = \{(HHHHH), (THHHH), (HTHHH), (HHTHH),$$
$$(HHHTH), (HHHHT), (TT), (HTT), (HHTT), (THTT),$$
$$(HHHTT), (THHHT), (HTHHT), (HHTHT), (HHTTH),$$
$$(THHTH), (HTHTH), (HTTHH), (THTHH), (TTHHH)\}$$

Outcomes in this sample space is not equally likely. For example $P(\{(TT)\}) = \frac{1}{4}$ and $P(\{(HTT)\}) = \frac{1}{8}$. The total probability is

$$P(S) = P(\{(TT)\}) + P(\{(HTT)\})$$
$$+ 2P(\{(HHTT)\}) + 16P(\{(HHHTT)\})$$
$$= \frac{1}{4} + \frac{1}{8} + 2\frac{1}{16} + 16\frac{1}{32}$$
$$= 1$$

as expected.

EXAMPLE 13.

In the game of poker, find $P(\text{four of a kind})$, $P(\text{full house})$, $P(\text{a pair})$, $P(\text{two pairs})$, and $P(\text{straight})$.

As mentioned before, the size of the sample space is $|S| = \binom{52}{5} = 2598960$. We need to find how many possible sets of four-of-a-kind $|E|$. There are 13 different values for the first four cards (there are four suit options and each of the four cards takes a distinct suit but we don't count permutations here) and once the four cards are out, there are 48 (12 choices for value and 4 choices for suit) options for the last card. So $|E| = 13(48) = 624$. Therefore, $P(\text{four of kind}) = \frac{|E|}{|S|} = 0.00024$.

There are 13 possible card values for the first three cards. They can take any suit and there are $\binom{4}{3} = 4$ different suit combinations for the first three cards. For the remaining two cards, there are 12 possible values and $\binom{4}{2} = 6$ different suit combinations. So $|E| = 3744$ and the probability is therefore $P(\text{full house}) = 0.00144$.

---

[1]$\mathcal{P}$ is power set

There are $\binom{13}{1}$ ways to pick a value for a pair. There are $\binom{4}{2}$ different suit combinations for a pair. There are $\binom{12}{3}$ ways to pick 3 distinct values for the rest of the 3 cards and each card has $\binom{4}{1}$ different suit combinations. The answer is $\frac{\binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3}{\binom{52}{5}} = 0.4226$.

There are $\binom{13}{2}$ ways to pick values for two pairs. There are $\binom{4}{2}$ different suit combinations for each pair. There are $\binom{11}{1}$ ways to pick a distinct value for the last card and has $\binom{4}{1}$ different suit combinations. The answer is $\frac{\binom{13}{2}\binom{4}{2}^2\binom{11}{1}\binom{4}{1}}{\binom{52}{5}} = 0.04754$.

There are 10 possible values for the initial card. There are 4 suit options for each of the five cards, so there are 4^5 suit combinations but to exclude straight flushes we subtract it by 4. This gives $|E| = 10(4^5 - 4) = 10200$, so $P(\text{straight}) = 0.00392$.

EXAMPLE 14.

There are 7 marbles in a bag, 4 red and 3 blue. You draw marbles one at a time without replacement. What is the probability that the fifth marble is blue?

One way to do this problem is to try labelling each marble from 1 to 7 which represents the order of drawing out. So the probability that the blue marble gets 5 is $\frac{3}{7}$.

EXAMPLE 15.

You roll five dice, find probabilities for (a) one pair ($aabcd$) (b) two pairs ($aabbc$) (c) full house ($aaabb$).

(a) There are $\binom{6}{4}$ ways to pick values for the set $\{a, b, c, d\}$. There are $\frac{5!}{2!}$ ways to arrange $aabcd$. There are 4 ways to select the value of a pair from the set $\{a, b, c, d\}$. There are 6^5 permutations in total. The final answer is $\frac{\binom{6}{4}\frac{5!}{2!}4}{6^5} = 0.46296$.

(b) Caring about order, there are $\binom{6}{2}$ ways to pick values for the pairs $\{a, b\}$ and $\binom{4}{1}$ ways to pick the value of c. There are $\frac{5!}{2!2!}$ ways to arrange $aabbc$. There are 6^5 permutations in total. The final answer is $\frac{\binom{6}{2}\binom{4}{1}\frac{5!}{2!2!}}{6^5} = 0.23148$.

(c) There are $\binom{6}{1}$ ways to pick values for a and $\binom{5}{1}$ ways to pick values for b. There are $\frac{5!}{3!2!}$ ways to arrange $aaabb$. The final answer is $\frac{\binom{6}{1}\binom{5}{1}\frac{5!}{3!2!}}{6^5} = 0.03858$.

Note on counting: $\binom{6}{1}\binom{5}{1}$ is not equal to $\binom{6}{2}$. The former is used if order matters otherwise the latter is used. In fact, $\binom{6}{1}\binom{5}{1} = 2\binom{6}{2} = {}_6P_2$. Imagine you have a bag of marbles labelled $1, 2, \ldots, 6$, $\binom{6}{2}$ counts the number of two marble combinations while $\binom{6}{1}\binom{5}{1}$ counts the number of ways to pick the first marble and paint it red then pick the second marble.

EXAMPLE 16.

A box contains 15 lego bricks: 3 red bricks, 3 green bricks, 3 blue bricks, 3 yellow bricks and 3 purple bricks. You draw 5 bricks from the box at random without replacement. What is the probability of getting 5 distinct colours?

There are $15 \cdot 14 \cdot 13 \cdot 12 \cdot 11$ different ordered selections (15 options for the first brick,

14 options for the second brick and so on). Now, there are 15 options for the first brick which can be in any colour. Then there are 12 options for the second brick (we exclude all bricks with the same colour as the first brick). There are 9 options for the third brick (we exclude all bricks with the same colour as the first and second bricks) and so on. The probability is

$$\frac{15}{15}\frac{12}{14}\frac{9}{13}\frac{6}{12}\frac{3}{11} = \frac{81}{1001}$$

EXAMPLE 17.

You have 13 cards, 6 red (R) and 7 black (B). You shuffle them and place them face up in two rows. The top row has 5 cards, the bottom row 8. If you do this many times, what number of red cards will the top row have most often?

The total number of arrangements is $\frac{13!}{6!7!} = 1716$.

To determine the most likely number of red cards, we count the number of arrangements when the top row has $0, 1, 2, 3, 4, 5$ red cards. For example, to count the number of arrangements for the top row with 2 red cards, $\binom{5}{2} = \frac{5!}{2!3!} = 10$. To get the total number of arrangements we multiply the results, for example, the total number of arrangements with 3 red cards at the top is $10(56) = 560$. After that, we calculate probabilities and compare.

no. red at top	0	1	2	3	4	5
no. arrangements for top	1	5	10	10	5	1
no. red at bottom	6	5	4	3	2	1
no. arrangements for bottom	28	56	70	56	28	8
total no. of arrangements	28	280	700	560	140	8
probability	0.016	0.163	0.408	0.326	0.082	0.005

From the table, we see that the top row will most likely have 2 red cards.

Also notice that $28 + 280 + 700 + 560 + 140 + 8 = 1716$.

If we were to compute the average number of red cards in the top row, we use the following relation: $\sum_{n=0}^{5} nP_n$ where P_n is the probability of getting n red cards in the top row. This gives:

$$0(0.016) + 1(0.163) + 2(0.408) + 3(0.326) + 4(0.082) + 5(0.005) = 2.31$$

This playing card example has applications in statistical mechanics for modelling two-state systems.

EXAMPLE 18.

You have a standard deck of 52 cards which consists of 26 red cards and 26 black cards. The deck is shuffled well, what is the probability that the first five cards contain 3 red cards and 2 black cards? What is the probability if the deck is large enough?

We count the number of ways to draw 3 red cards from the set of 26 red cards: $\binom{26}{3}$ and the number of ways to draw 2 black cards from the set of 26 black cards: $\binom{26}{2}$. Then multiply the result and divide it by the number of ways to draw 5 cards from the deck: $\binom{52}{5}$. This gives $P(3R, 2B) = \frac{\binom{26}{3}\binom{26}{2}}{\binom{52}{5}}$.

Alternative approach: Count the number of combinations of a 5 card hand that contains 3 red and 2 black: $\binom{5}{3}$. Then count the number of ways to rearrange the remaining cards in the deck: $\binom{52-5}{26-3}$. Finally count the number of ways to arrange the deck: $\binom{52}{26}$. This gives $P(3R, 2B) = \frac{\binom{5}{3}\binom{52-5}{26-3}}{\binom{52}{26}}$.

To approximate the probability when the deck is large, we rewrite things a bit:

$$
\begin{aligned}
P(3R, 2B) &= \frac{\binom{26}{3}\binom{26}{2}}{\binom{52}{5}} \\
&= \frac{\frac{26!}{3!23!}\frac{26!}{2!24!}}{\frac{52!}{5!47!}} \\
&= \frac{5!}{3!2!}\frac{\frac{47!}{23!24!}}{\frac{52!}{26!26!}} \\
&= \frac{5!}{3!2!}\frac{(52-5)!26!26!}{52!(26-3)!(26-2)!} \\
&= \frac{5!}{3!2!}\frac{26(26-1)(26-2)26(26-1)}{52(52-1)(52-2)(52-3)(52-4)} \\
&\approx \frac{5!}{3!2!}\frac{26^5}{56^5} \\
&\approx \frac{5!}{3!2!}\frac{1}{2^5}
\end{aligned}
$$

We conclude that the suitable approximation is $\frac{\binom{5}{3}}{2^5}$. This gets better when the deck is larger.

This example illustrates the concept of heat bath. The larger the deck is, taking out a few cards will have almost no effect on the probability.
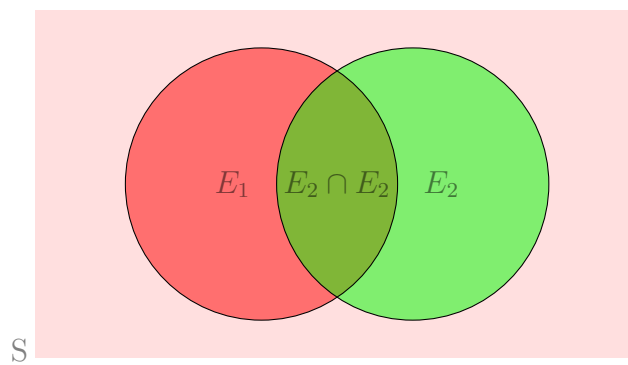
2.3 Useful Probability Relations

Probability has the following useful properties:

 i For a single event E, $P(S) = P(E) + P(E^c) = 1$. E and E^c is always disjoint.
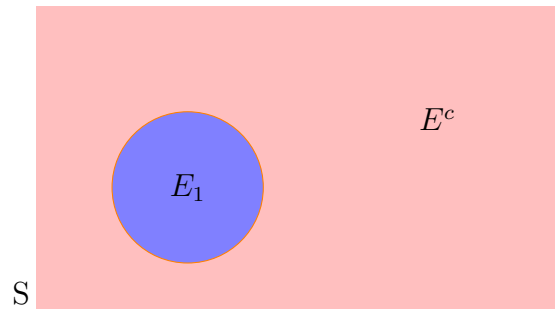
Figure 3: Event E and its complement E^c

ii For two non-disjoint events E_1 and E_2, probability can be calculated using inclusion-exclusion principle: $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.

Figure 4: Cartoon illustrating non-disjoint events E_1 and E_2

For disjoint events, $P(E_1 \cap E_2) = 0$ since $E_1 \cap E_2 = \emptyset$.

For three events, it becomes:

$$
\begin{aligned}
P(E_1 \cup E_2 \cup E_3) = {} & P(E_1) + P(E_2) + P(E_3) \\
& - P(E_1 \cap E_2) - P(E_2 \cap E_3) - P(E_1 \cap E_3) \\
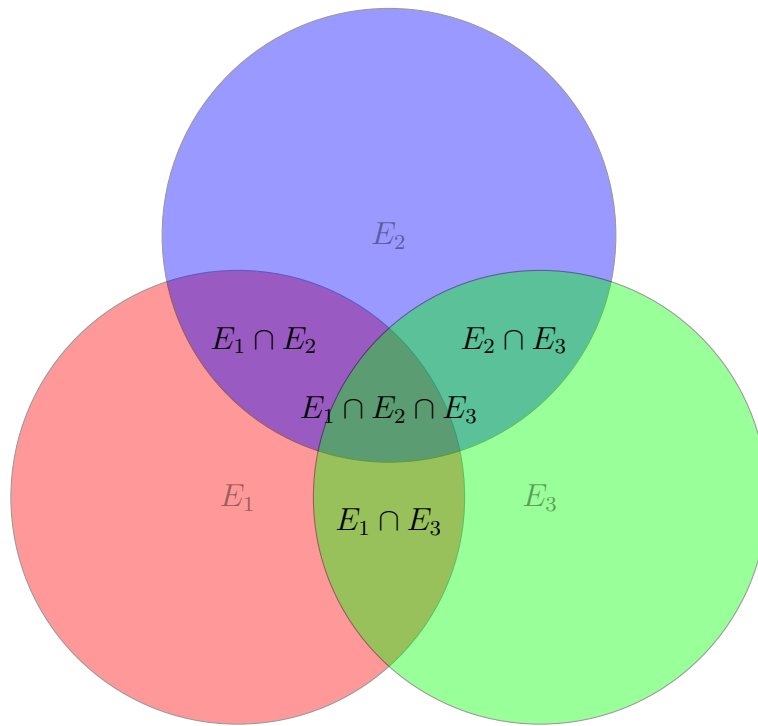& + P(E_1 \cap E_2 \cap E_3)
\end{aligned}
$$

Figure 5: Three non-disjoint events E_1, E_2 and E_3

For more than three events, the inclusion-exclusion principle can be generalized to:

$$
\begin{aligned}
P\left(\bigcup_{n=1}^{N} E_n\right) = & \sum_i P(E_i) \\
& - \sum_{i<j} P(E_i \cap E_j) \\
& + \sum_{i<j<k} P(E_i \cap E_j \cap E_k) \\
& - \sum_{i<j<k<l} P(E_i \cap E_j \cap E_k \cap E_l) + \cdots + (-1)^{n-1} P\left(\bigcap_{n=1}^{N} E_n\right)
\end{aligned}
$$

EXAMPLE 19.

Birthday problem: There are n people in the room, what is the probability that at least two of them share the same birthday?

The sample space S for this experiment consists of sequences of n birthdays: $(B_1, B_2, \ldots, B_n) \in S$. The size of the sample space is $|S| = 365^n$, ignoring leap years. The event E is the set of sequences that contains at least two matching birthdays, like for example $E = \{\ldots (09/11, \ldots, 09/11) \ldots\}$.

We first calculate the probability that all n birthdays are different. This is the complement of the event E, denoted by E^c. There are 365 unique choices for the first person's birthday and this leaves with 364 choices for the second person's birthday and so on. The probability is:

$$P(\text{all different}) = \frac{|E^c|}{|S|} = \frac{365 \cdot 364 \cdot 363 \cdots (365 - n + 1)}{365^n} = \frac{_{365}\mathrm{P}_n}{365^n}$$

The probability that at least two people share the same birthday is therefore:

$$P(\text{at least two same bday}) = 1 - P(\text{all different}) = 1 - \frac{_{365}\mathrm{P}_n}{365^n}$$

To illustrate the result, suppose $n = 23$, the probability is $P(\text{at least two same bday}) = 0.5073$.

EXAMPLE 20.

You flip two coins. Find the probability that either the first or the second coin fall heads.

The possible outcomes S for this experiment are: $S = \{(H, H), (H, T), (T, H), (T, T)\}$. Let E_1 be the event that the first coin falls heads: $E_1 = \{(H, H), (H, T)\}$ and E_2 be the event that the second coin falls heads: $E_2 = \{(H, H), (T, H)\}$. By the inclusion-exclusion principle:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$
$$= \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$$

It can be computed directly as well: $P(E_1 \cup E_2) = P(\{(H, H), (H, T), (T, H)\}) = \frac{3}{4}$.

EXAMPLE 21.

Find the probability that the random permutation of 1112345 does not have adjacent identical digits.

We label 1s to distinguish them: $1_a 1_b 1_c 2345$. The sample space S is all permutations of $1_a 1_b 1_c 2345$ and the size is $|S| = 7! = 5040$ (remember that 1s are distinct from each other!). Let E_{ab}, E_{bc}, E_{ac} denote events where "1_a and 1_b are adjacent," "1_b and 1_c are adjacent" and "1_a and 1_c are adjacent" respectively. The probability that there are no adjacent ones can be calculated using the inclusion-exclusion principle:

$$\begin{aligned}
P(\text{no adjacent 1s}) &= 1 - P(E_{ab} \text{ or } E_{bc} \text{ or } E_{ac}) \\
&= 1 - P(E_{ab}) - P(E_{bc}) - P(E_{ac}) \\
&\quad + P(E_{ab} \text{ and } E_{bc}) + P(E_{ab} \text{ and } E_{bc}) \\
&\quad + P(E_{ac} \text{ and } E_{bc}) \\
&\quad - P(E_{ab} \text{ and } E_{bc} \text{ and } E_{ac})
\end{aligned}$$

By symmetry and the fact that $1_a, 1_b, 1_c$ cannot *all* be adjacent to each other:

$$P(\text{no adjacent 1s}) = 1 - 3P(E_{ab}) + 3P(E_{ab} \text{ and } E_{bc})$$

$P(E_{ab}) = \frac{2(6!)}{7!} = \frac{2}{7}$: if 1_a and 1_b are adjacent, we can treat them as one symbol $[1_a 1_b]$, so there are 6! ways to order the 6 symbols we now have, and then 2 ways to choose the order of 1_a and 1_b. Similarly, $P(E_{ab} \text{ and } E_{bc})$ can be calculated by treating $[1_a 1_b 1_c]$ as a

symbol. There are two valid permutations where $[1_a 1_b]$ and $[1_b 1_c]$ combinations exist, namely $[1_a 1_b 1_c]$ and $[1_c 1_b 1_a]$. This gives $\frac{2(5!)}{7!} = \frac{1}{21}$.

The final answer is:

$$P(\text{no adjacent 1s}) = 1 - 3\frac{2}{7} + 3\frac{1}{21} = \frac{2}{7}$$

EXAMPLE 22.

n people put their names in a hat. The hat is shaken and n people draw names out of the hat (secret Santa). What is the probability that nobody gets their own name?

We label each person from 1 to n. The sample space S is all permutations of $\{1, 2, \ldots, n\}$ and its size is $|S| = n!$. Let E_i denote an event where person i gets their own name: $E_i = \{\text{permutations where person } i \text{ gets their own name.}\}$. We first calculate $P(E_1 \cup E_2 \cup \cdots \cup E_n)$.

First, we calculate $P(E_i)$. The number of events where person 1 gets his own name $|E_1|$ can be calculated by placing person 1 on position 1 from the start and counting the number of ways to arrange the remaining $n - 1$ people on positions $2, 3, \ldots n$ which is $(n-1)!$. The probability that person 1 gets his own name is $P(E_1) = \frac{(n-1)!}{n!} = \frac{1}{n}$. By symmetry, $P(E_i) = P(E_1)$ for $i = 1, 2, \ldots n$.

Next we calculate $P(E_i \cap E_j)$. $|E_1 \cap E_2|$ is the number of events where both person 1 and person 2 gets their own names. The calculation process is similar: place persons 1 and 2 at positions 1 and 2 respectively and calculate the number of ways to arrange the remaining $n - 2$ people on positions $3, \ldots n$ which is $(n - 2)!$. So, $P(E_1 \cap E_2) = \frac{(n-2)!}{n!}$ and by symmetry, $P(E_i \cap E_j) = P(E_1 \cap E_2)$ for any $i < j$. There are $\binom{n}{2}$ pairs of i, j.

We can generalize that the probability that k people get their own names is $\frac{(n-k)!}{n!}$ and there are $\binom{n}{k}$ combinations of k people. Now by the inclusion-exclusion principle:

$$P\left(\bigcup_{n=1}^{N} E_n\right) = nP(E_1) - \binom{n}{2}P(E_1 \cap E_2) + \binom{n}{3}P(E_1 \cap E_2 \cap E_3)$$

$$+ \cdots \pm (-1)^{n-1}P\left(\bigcap_{n=1}^{N} E_n\right)$$

$$= n\frac{1}{n} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} - \cdots + \frac{(-1)^{n-1}}{n!}$$

$$= \sum_{k=1}^{N}(-1)^{k-1}\binom{n}{k}\frac{(n-k)!}{n!}$$

$$= \sum_{k=1}^{N}(-1)^{k-1}\frac{1}{k!} = 1 - e^{-1}$$

Therefore the probability that nobody draws their own name is $1 - (1 - e^{-1}) = e^{-1} = 0.368$.

2.4 Conditional Probability and Independent Events

Suppose there are two events E and F. Assume that $P(F) > 0$. The probability that E happens given that F happened is denoted by $P(E|F)$. Imagine E and F lie on s sample space S.
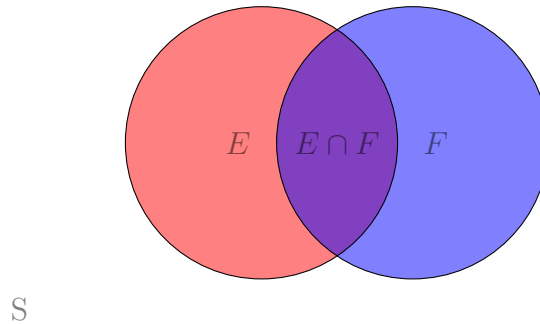

Figure 6: Events E and F

Recall that the probability of $P(E)$ is proportional to the probability of the sample space $P(S) = 1$. When event F happens, F will become a new sample space as illustrated below.
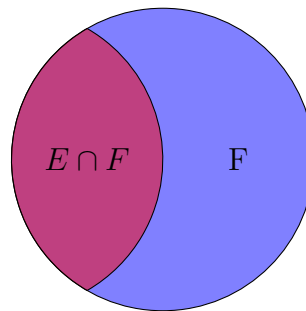

Figure 7: New sample space F

With this new sample space, we can be sure that everything happens on F. $E|F$ indicates both events E and F happen, so we have $E \cap F$ lie in F. We can now define $P(E|F)$ as a ratio of $E \cap F$ to F, assuming that every outcome in F is equally likely:

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

Dividing both numerator and denominator by $|S|$ gives:

$$P(E|F) = \frac{\frac{|E \cap F|}{|S|}}{\frac{|F|}{|S|}} = \frac{P(E \cap F)}{P(F)}$$

We derived the formula for conditional probability. It satisfies all probability axioms. As a corollary, this formula is also written in this way:

$$P(E \cap F) = P(F)P(E|F)$$

This is intuitive: "probability for both E and F happen is the probability that F happens times the probability that E happens *given* that F happened."

Two events E and F are independent if finding out one event does not affect the probability of another, that is $P(E|F) = P(E)$. This means uncovering the existence of F does not affect the probability of E happening. We can test whether events are independent or not using the below relation:

$$P(E \cap F) = P(F)P(E)$$

Note that disjoint and independent are two different things. Events being disjoint means they are mutually exclusive. Here are some examples of (non-)disjoint and (in)dependent events.

1. **Disjoint but dependent events**

   Consider an experiment of rolling a die. Define events E to be "roll 3" and F to be "roll 4." These events are disjoint because you can't roll both 3 and 4 at the same time using a single dice and this makes $P(E \cap F) = 0$. It implies that events are not independent (ie. dependent) because of $P(E \cap F) \neq P(E)P(F)$. Also since $P(E \cap F) = 0$, $P(E|F) = 0$ which makes sense in this context.

   We can conclude that disjoint events cannot be independent. There are exceptions, for example, consider the game of darts. Define events E "dart land on target" and F "dart land on the edge of the board." $P(E \cap F) = 0$ since a dart land at both places same time but $P(E) = 0$ and $P(F) = 0$ since there are infinitely many points but it is still possible to land on one of them. Events E and F are disjoint but independent events.

2. **Independent but not disjoint events**

   Consider the dice experiment again but with two dice this time. Define events E "the sum of the face values is 7" and F the face value of the first die is 2. They are not disjoint because there is a possible outcome where the sum of the face value is 7 and the face value of the first die is 2: $\{2, 5\}$. E and F are independent because $P(E \cap F) = \frac{1}{36}$ and $P(E)P(F) = \frac{1}{36}$. But if we redefine E to be "the sum of the face values is 6," the events will become dependent because the probability depends on whether we rolled 6 on the first roll or not (we must roll numbers 1 to 5 to get some chance that the face values totalling 6).

3. **Dependent but not disjoint events**

   We discussed an example of dependent but not disjoint events previously but here is another one: In an experiment where you draw two balls from a bag containing 5 red balls and 3 green balls without replacement, define events E "draw a red ball on the first draw" and F "two balls are red." They are not disjoint because there is an outcome where E and F happen at the same time: $\{R, R\}$. They are dependent (ie. not independent) because $P(E \cap F) = P(E)P(F|E) = \frac{5}{8}\frac{4}{7} = \frac{5}{14}$ and $P(E)P(F) = \frac{5}{8}\frac{5}{14} = \frac{25}{112}$.

For more than two events, events $E_1, E_2, \ldots, E_n$ are independent iff

$$P\left(\bigcup_i^k E_{i'}\right) = \prod_i^k P(E_{i'})$$

for every subset $E_{1'}, E_{2'}, \ldots . E_{k'}$ of size $k \leq n$.

If events A, B, C, D are dependent:

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

EXAMPLE 23.

You draw two cards from a standard deck. Define events E "draw hearts as the first card" and F "draw hearts as the second card." Find $P(\text{draw two hearts})$.

We have $P(E) = P(F) = \frac{13}{52}$ and

$$P(E \cap F) = P(E)P(F|E) = \frac{13}{52}\frac{12}{51} = \frac{1}{17}$$

or alternatively, $P(E \cap F) = \frac{\binom{13}{2}}{\binom{52}{2}} = \frac{1}{17}$.

The events E and F are dependent because $P(E \cap F) \neq P(E)P(F)$.

EXAMPLE 24.

You draw a card from a standard deck. Define events E "draw red card" and F "draw ace." Find $P(\text{red ace})$.

We find that $P(E) = \frac{1}{2}$ and $P(F) = \frac{4}{52} = \frac{1}{13}$. To find $P(\text{red ace})$, we calculate $P(E \cap F)$:

$$P(E \cap F) = P(F)P(E|F) = \frac{1}{13}\frac{1}{2} = \frac{1}{26}$$

Notice that $P(E|F) = P(E)$ which makes events E and F independent.

EXAMPLE 25.

You toss two coins. Define events E "first coin heads," F "second coin heads" and G "both heads or both tails." Are these three events mutually independent?

The sample space is $S = \{HH, HT, TH, TT\}$. We have $P(E) = \frac{1}{2}$, $P(F) = \frac{1}{2}$ and $P(G) = \frac{1}{2}$. Any of these two events are independent, for example, $P(E \cap F) = \frac{1}{4}$ which is equal to $P(E)P(F) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$. However, $P(E \cap F \cap G) = \frac{1}{4}$ which is not equal to $P(E)P(F)P(G) = \frac{1}{8}$, so the three events combined are not independent.

EXAMPLE 26.

You roll two dice blindfolded. You are told that the sum is not greater than 3, what is the probability that both dice have the same face value?

Let E be the event when both dice have the same face value and F be the event when the sum is not greater than 3. Since they both share the $\{1, 1\}$ outcome, the events are not disjoint.

We wish to calculate $P(E|F)$. We have $P(E) = \frac{6}{36} = \frac{1}{6}$ and $P(F) = \frac{3}{36} = \frac{1}{12}$ (the possible outcomes are $\{1, 1\}$, $\{1, 2\}$ and $\{2, 1\}$). The probability to get both events is $P(E \cap F) = \frac{1}{36}$. This makes events E and F dependent since $P(E \cap F) \neq P(E)P(F)$. Therefore $P(E|F)$ is given by:

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{\frac{1}{36}}{\frac{1}{12}} = \frac{1}{3}$$

2.5   Law of Total Probability

Let $F_1, F_2, \ldots, F_n$ denote partitions of sample space S in such way that

$$S = \bigcup_i^n F_i$$

There is also an event E in S. We want to find the relationship between E and $F_1, F_2, \ldots, F_n$.
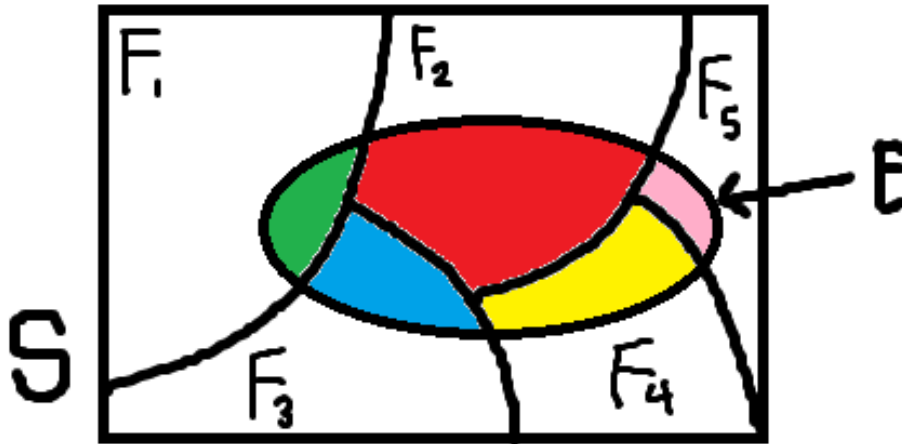

Figure 8: Partitions of S

From the diagram, we see that E is expressed as disjoint unions (coloured regions) of $E \cap F_i$, so

$$P(E) = \sum_i P(E \cap F_i) = \sum_i P(F_i)P(E|F_i)$$

This is the law of total probability. It expresses $P(E)$ in terms of weighted averages of $P(E|F_i)$s. For two events E and F (see figure 6), E can be expressed as $(E \cap F) \cup (E \cap F^c)$ and the law becomes:

$$\begin{aligned}
P(E) &= P(E \cap F) + P(E \cap F^c) \\
&= P(F)P(E|F) + P(F^c)P(E|F^c) \\
&= P(F)P(E|F) + (1 - P(F))P(E|F^c)
\end{aligned}$$

EXAMPLE 27.

You choose a random number from 1 to 10 and then you throw that many dice. Define events E "sum of the face values is 3" and F_i "choose i." Find $P(E)$.

We have $P(F_i) = \frac{1}{10}$ for all i and $P(E|F_1) = \frac{1}{6}$ (given that we selected 1, we need to roll 3 to get the desired sum), $P(E|F_2) = \frac{2}{36}$ (there are two outcomes that give the desired sum: $\{1, 2\}$ and $\{2, 1\}$) and $P(E_3) = \frac{1}{6^3}$ (the only possible outcome is $\{1, 1, 1\}$). For $i > 3$, achieving the desired sum is impossible at this point so $P(E|F_i) = 0$. Now $P(E)$ can be computed as follows:

$$P(E) = \sum_{i=1}^{10} P(F_i)P(E|F_i)$$
$$= P(F_1)P(E|F_1) + P(F_2)P(E|F_2) + P(F_3)P(E|F_3)$$
$$+ P(F_4)P(E|F_4) + \cdots + P(F_{10})P(E|F_{10})$$
$$= \frac{1}{10}\frac{1}{6} + \frac{1}{10}\frac{2}{36} + \frac{1}{10}\frac{1}{6^3} + \frac{1}{10}0 + \cdots + 0$$
$$= \frac{49}{2160}$$

2.6 Bayes' Theorem

For events E and F, we found that

$$P(E \cap F) = P(F)P(E|F)$$

but $P(E \cap F)$ can also be calculated in terms of $P(E)$ which gives:

$$P(E \cap F) = P(E)P(F|E)$$

Combining these equations gives:

$$\boxed{P(F|E) = \frac{P(E|F)P(F)}{P(E)}}$$

The formula expressing the relationship between $P(E|F)$ and $P(F|E)$ is known as Bayes' theorem. This powerful relationship has many real-world applications like detecting false positives in medical diagnosis. Once you know $P(E|F)$, $P(E)$ and $P(F)$, then it will be easy to find $P(F|E)$.
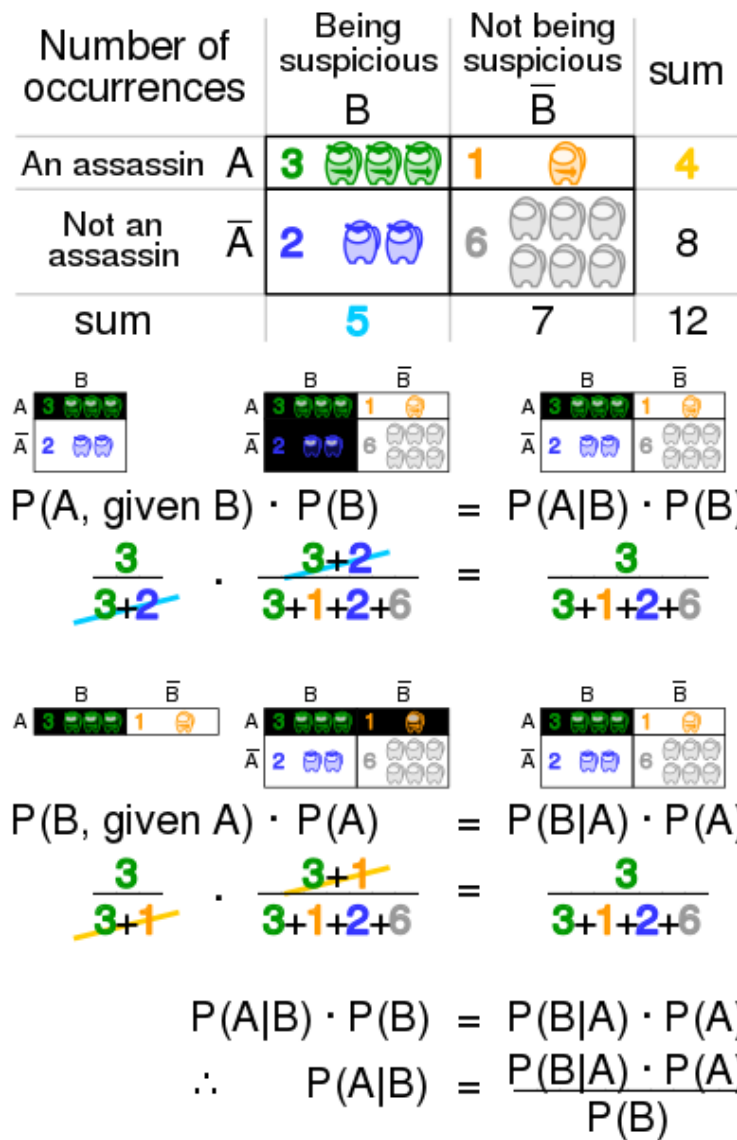
Number of occurrences	Being suspicious B	Not being suspicious $\overline{B}$	sum
An assassin A	3	1	4
Not an assassin $\overline{A}$	2	6	8
sum	5	7	12

$$P(A, \text{given } B) \cdot P(B) = P(A|B) \cdot P(B)$$

$$\frac{3}{3+2} \cdot \frac{3+2}{3+1+2+6} = \frac{3}{3+1+2+6}$$

$$P(B, \text{given } A) \cdot P(A) = P(B|A) \cdot P(A)$$

$$\frac{3}{3+1} \cdot \frac{3+1}{3+1+2+6} = \frac{3}{3+1+2+6}$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$\therefore \quad P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Figure 9: Geometric interpretation of Bayes' theorem using Amogus

If we apply the law of total probability we derived from the previous section to $P(E)$:

$$P(F|E) = \frac{P(E|F)P(F)}{P(F)P(E|F) + P(F^c)P(E|F^c)}$$

For multiple events $F_1, F_2, \ldots, F_n$, it becomes:

$$P(F_j|E) = \frac{P(E|F_j)P(F)}{\sum_i P(E|F_i)P(F_i)}$$

This is useful if we want to determine which one of F_j occurred given that event E happened.

EXAMPLE 28.

Continuing from the last example, what is the probability that you selected 2 given that the sum is 3?

$$P(F_2|E) = \frac{P(E|F_2)P(F_2)}{P(F_1)P(E|F_1) + P(F_2)P(E|F_2) + P(F_3)P(E|F_3)}$$

$$= \frac{\frac{2}{36}\frac{1}{10}}{\frac{1}{10}\frac{1}{6} + \frac{1}{10}\frac{2}{36} + \frac{1}{10}\frac{1}{6^3}} = \frac{12}{49}$$

EXAMPLE 29.

A medical test is carried out to test whether a person has cancer or not. The test is 95% effective at detecting whether the person has cancer or not when it is actually present. However, there is 1% percent of chance of false positive. False positives happen when the test gives a positive result even though the person is perfectly healthy. False negatives happen when the test gives a negative result even though the person actually has cancer. We know that about one in a hundred people have cancer. Find the probability that a person has cancer given that the test gives a positive result.

Define events E "test results in positive," H "person is healthy" and C "person has cancer."

From the description, we know that $P(C) = \frac{1}{100}$, $P(E|C) = \frac{95}{100}$ and $P(E|H) = \frac{1}{100}$. We can infer that $P(H) = 1 - P(C) = \frac{99}{100}$, $P(\text{false positive}) = P(E|H) = \frac{1}{100}$, and $P(\text{false negative}) = 1 - P(E|C) = \frac{5}{100}$.

We want to find $P(C|E)$ and it is calculated as follows:

$$\begin{aligned}
P(C|E) &= \frac{P(E|C)P(C)}{P(E)} \\
&= \frac{P(E|C)P(C)}{P(C)P(E|C) + P(H)P(E|H)} \\
&= \frac{\frac{95}{100}\frac{1}{100}}{\frac{1}{100}\frac{95}{100} + \frac{99}{100}\frac{1}{100}} \approx 0.49
\end{aligned}$$

EXAMPLE 30.

Monty Hall problem: The host puts a prize in one of the three doors. You choose door 1. The host then reveals door 2 which contains no prize. Should you make a switch to door 3?

Let E be an event when the host opens door 2 which contains no prize (he can also open other doors not just this door) and F_i be an event when the prize is located in door i.

Since we don't know which door the prize is in, $P(F_1) = P(F_2) = P(F_3) = \frac{1}{3}$. The probability that the host will open door 1 is $P(E|F_1) = \frac{1}{2}$ because he can either open door 2 or door 3 in this case. $P(E|F_2) = 0$ because there is no way he would reveal the prize door. $P(E|F_3) = 1$ because door 2 is the only door he can open in this case.

We want to find $P(F_1|E)$ and $P(F_3|E)$. By Bayes' theorem, we get

$$P(F_1|E) = \frac{P(E|F_1)P(F_1)}{\sum_{j=1}^{3} P(E|F_j)P(F_j)} = \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + 0\frac{1}{3} + 1\frac{1}{3}} = \frac{1}{3}$$

and

$$P(F_2|E) = \frac{P(E|F_2)P(F_2)}{\sum_{j=1}^{3} P(E|F_j)P(F_j)} = \frac{1\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + 0\frac{1}{3} + 1\frac{1}{3}} = \frac{2}{3}$$

We definitely should make a switch!

$$P(F_1|E) = \frac{P(E|F_1)P(F_1)}{\sum_{j=1}^{3} P(E|F_j)P(F_j)} = \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + 0\frac{1}{3} + 1\frac{1}{3}} = \frac{1}{3}$$

3 Random Variables

3.1 Introduction

A random variable is a function that maps elements from a sample space to any real number: $X : S \to \mathbb{R}$. It is usually denoted by X, Y, Z, N. Let s be a member of a sample space S, if s is picked up by probability function P then random variable $X(s)$ returns a real number. A random variable can either be discrete or continuous.

A discrete random variable only takes finite (or countably infinite) m values: $\{x_1, x_2, \ldots, x_m\}$. The sample space is broken into discrete partitions.
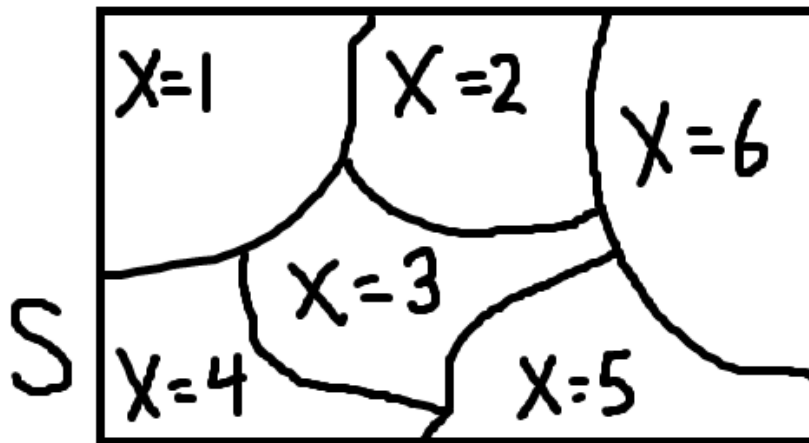
Figure 10: Example of a discrete R.V.

On the other hand, a continuous random variable can take infinitely many values. One such example is a coordinate on the x-axis which is represented by a real number: $X(x) = x$.

The probability when a random variable takes a certain value is denoted by $P(X = a)$ which is the shortcut for writing $P(\{s \in S \mid X(s) = a\})$. Ranges are also allowed: $P(X \leq a) = P(\{s \in S \mid X(s) \leq a\})$.

EXAMPLE 31.

You roll two dice. Find probabilities of each sum. What is the probability of rolling 30 or π?

The sample space for two dice rolls is $S = \{(a, b) \mid a, b \in \{1, 2, \ldots, 6\}\}$. The random variable relevant to this problem is $X((a, b)) = a + b$. Below are probabilities for each possible sum.

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

$$P(X = 3) = P((1,2),(2,1)) = \frac{2}{36}$$

$$P(X = 4) = P((1,3),(2,2),(3,1)) = \frac{3}{36}$$

$$P(X = 5) = P((1,4),(2,3),(3,2),(4,1)) = \frac{4}{36}$$

$$P(X = 6) = P((1,5),(2,4),(3,3),(4,2),(5,1)) = \frac{5}{36}$$

$$P(X = 7) = P((1,6),(2,5),(3,4),(4,3),(5,2),(6,1)) = \frac{6}{36}$$

$$P(X = 8) = P((2,6),(3,5),(4,4),(5,3),(6,2)) = \frac{5}{36}$$

$$P(X = 9) = P((3,6),(4,5),(5,4),(6,3)) = \frac{4}{36}$$

$$P(X = 10) = P((4,6),(5,5),(6,4)) = \frac{3}{36}$$

$$P(X = 11) = P((5,6),(6,5)) = \frac{2}{36}$$

$$P(X = 12) = P((6,6)) = \frac{1}{36}$$

Notice that the total probability adds up to one

$$\sum_{n=2}^{12} P(X = n) = 1$$

As for 30 and π, the probability is zero.

3.1.1 Commutative Distribution Function

We conclude this section with brief words on commutative distribution functions (CDF). CDF for a random variable X is defined as

$$F_X(a) = P(X \le a)$$

for all a values. In other words, $F_X(a)$ is the sum of probabilities from $-\infty$ to a.

The basic properties of $F_X(a)$ are:

  i $0 \le F_X(a) \le 1$

 ii $F_X(a_1) \le F_X(a_2)$ for $a_1 < a_2$

iii $\lim_{a \to -\infty} F_X(a) = 0$

 iv $\lim_{a \to \infty} F_X(a) = 1$

The last two properties indicate the function $F_X(a)$ is monotone increasing.

We can calculate probabilities from CDFs using the following relations:

  i $P(a < X \leq b) = F_X(b) - F_X(a)$

 ii $P(X > a) = 1 - F_X(a)$

Notice that the first property vaguely resembles the fundamental theorem of calculus. We will see later that CDFs are what we get when we integrate probability density functions (PDF).

3.2   Discrete Random Variables

Since discrete random variables take a finite number of values, their CDF $F_X(a)$ changes in discrete steps for each value and are constant between jumps (think of stairs). Before we go further, we need to discuss probability mass functions (PMF) first.

3.2.1   Probability Mass Function

PMFs $p_X(x)$ are like bar charts used for calculating probability involving discrete random variables. They are defined in such a way that

$$p_X(a) = P(X = a)$$

for a discrete random variable X which takes the following values: $\{x_1, x_2, \ldots, x_m\}$. They satisfy the following properties:

  i $p_X(x_i) \geq 0$ for all i

 ii $\sum_i^m p_X(x_i) = 1$

$F_X(a)$ for a discrete variable can be found using: $F_X(a) = P(X \leq a) = \sum_{x_i \leq a} p_X(x_i)$.

We either use $p_X(x)$ or $F_X(a)$ to specify a discrete random variable. We will take a look at some special discrete random variables.

3.2.2   Bernoulli Random Variable

Bernoulli random variable $X \sim \text{Bern}(p)$ is useful for experiments involving binary outcomes, like "pass" and "fail" or "positive" or "negative." X either takes 0 or 1 where 0 represents "failure" and 1 represents "success."
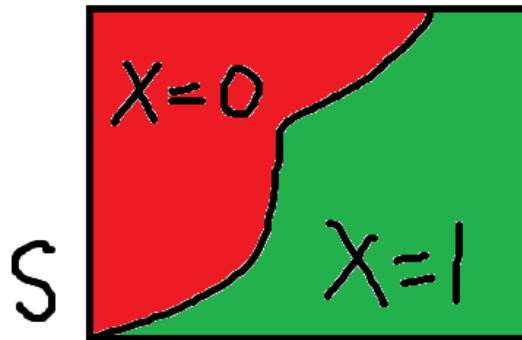

Figure 11: Bernoulli Random Variable

Suppose the probability for "success" is p, PMF for X is given by:

$$p_X(n) = P(X = n) = p^n(1-p)^{1-n}$$

where $p_X \in [0, 1]$ and $n = 0, 1$, or alternatively:

$$p_X(1) = P(X = 1) = p$$
$$p_X(0) = P(X = 0) = 1 - p$$

The CDF $F_X(a)$ is given by

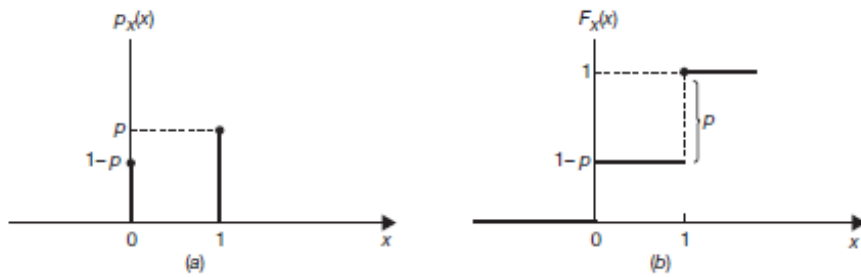$$F_X(a) = \begin{cases} 0 & a < 0 \\ 1 - p & 0 \le a < 1 \\ 1 & x \ge 1 \end{cases}$$

Figure 12: Bernoulli Distribution

3.2.3 Geometric Random Variable

Suppose you are carrying out some number of Bernoulli trials that has a success rate of q each time independently. You want to calculate the probability that after n tries you achieved your first success. The probability is calculated as follows:

$$P(X = n) = \underbrace{(1-q)(1-q)\cdots(1-q)}_{n-1 \text{ times}} q = (1-q)^{n-1}q$$

The suitable random variable for this problem is the geometric random variable $X \sim \text{Geom}(q)$. The possible values of n are $1, 2, \ldots$. Its PMF is given by

$$p_X(n) = P(X = n) = (1-q)^{n-1}q$$

where $p \in (0, 1]$ and $n = 1, 2, \ldots$. Its CDF is given by

$$F_X(a) = 1 - (1-q)^a$$

where $a = 1, 2, \ldots$.

Note that this random variable is still discrete even though it can take infinitely many values because they take countably infinite integer values.
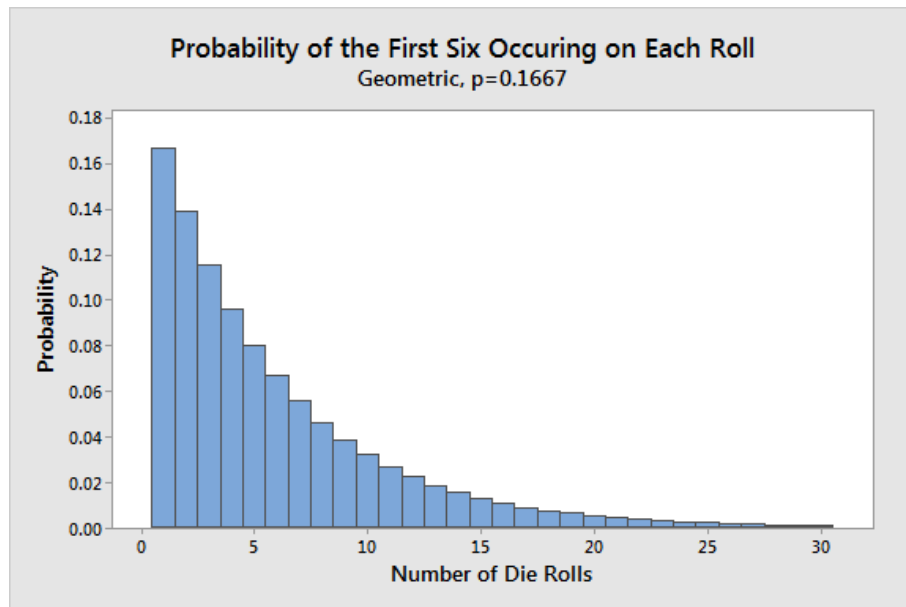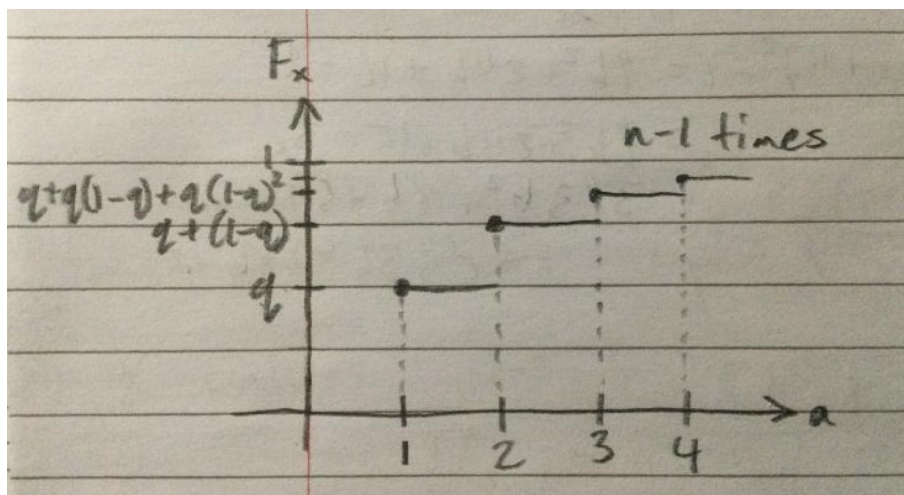
Figure 13: Example Geometric PMF

Figure 14: Geometric CDF

For a discrete random variable, $\sum_i p_X(x_i) = 1$ but for geometric random variable which is countably infinite,

$$\sum_{n=1}^{\infty} p_X(n) = \sum_{n=1}^{\infty} (1-q)^{n-1} q = 1$$

It can be proved using the formula for infinite geometric series:

$$\sum_{n=1}^{\infty} (1-q)^{n-1} q = q \sum_{n=1}^{\infty} (1-q)^{n-1}$$
$$= q \sum_{n=0}^{\infty} (1-q)^n$$
$$= q \frac{1}{1-(1-q)} = 1$$

A geometric random variable has something called *memoryless property*. The idea is that if an experiment failed n times then the probability that m additional failures will happen is the same as the probability that the experiment failed m times at the start. It indicates that the probability does not depend on past events. This property can be summarized by the below equation:

$$P(X > n + m \mid X > n) = P(X > m)$$

Explicit example: Consider an experiment of coin tosses and you happened to flip 5 tails in a row and with that information, the probability of getting an additional 2 tail is exactly the same as the probability of getting 2 tails.

Also, note that

$$P(X > n + m \mid X > n) = P(X > n + m)$$

is not always true unless events $X > n + m$ and $X > n$ are independent (ie. $P(X > n) = 1$).
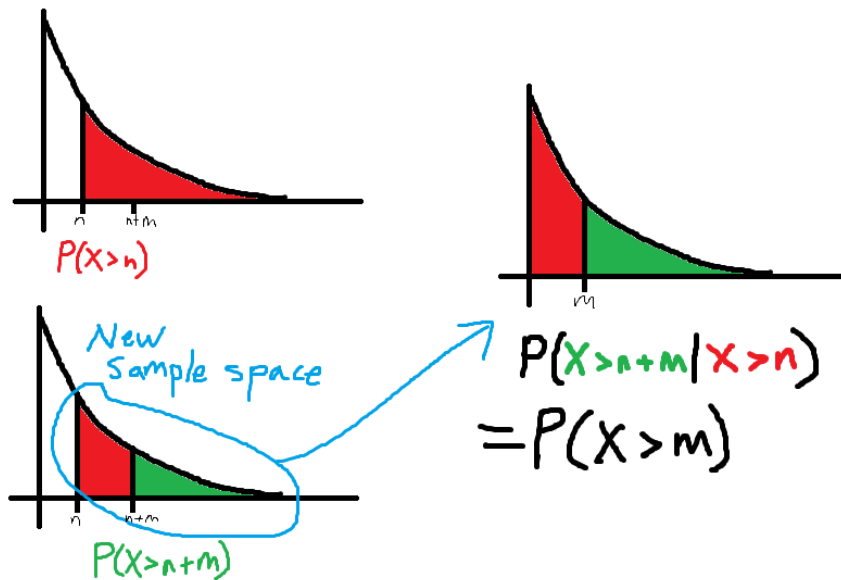

Figure 15: Visualization for Memoryless Property on PDF

To prove the memoryless property, consider

$$
\begin{aligned}
P(X > n) &= \sum_{a=n+1}^{\infty} p_X(a) \\
&= \sum_{a=n+1}^{\infty} (1-q)^{a-1} q \\
&= q(1-q)^n \sum_{a=0}^{\infty} (1-q)^a \\
&= (1-q)^n \sum_{a=0}^{\infty} (1-q)^a q
\end{aligned}
$$

Since the sum of the whole PMF is one,

$$= (1 - q)^n$$

Applying the formula for conditional probability,

$$P(X > n + m \mid X > n) = \frac{P(X > n + m \cap X > n)}{P(X > n)}$$
$$= \frac{(1 - q)^{n+m}}{(1 - q)^n}$$
$$= (1 - q)^m = P(X > m)$$

3.2.4   Binomial Random Variable

Binomial random variable $X \sim \text{Bin}(n, q)$ describes the number of successes if an experiment is repeated n times, each with success rate q. Each experiment must be independent of each other. Examples of such experiments like coin flips or any experiment with binary outcomes.

Its PMF is given by

$$p_X(k) = P(X = k) = \binom{n}{k} q^k (1 - q)^{n-k}$$

where $k = 0, 1, 2, \ldots, n$. This formula gives the probability of getting exactly k successes out of n trials.
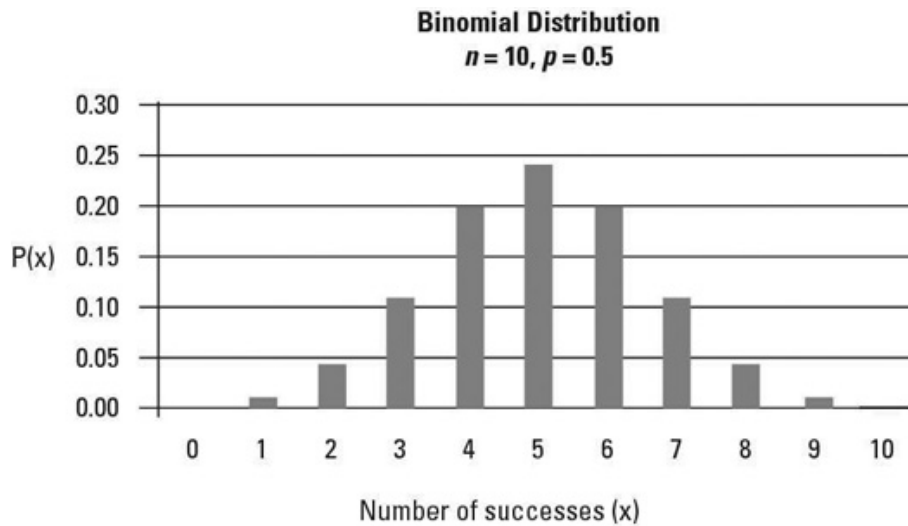

Figure 16: Example binomial PMF

The below example will build intuition on where the formula came from.

> EXAMPLE 32.
>
> Consider an unfair coin, let the probability of getting H on a single toss be 0.6 and the probability of getting T is 0.4. Find (1) the probability of obtaining the sequence $THHTTTHT$ (exactly, in this order) (2) the probability that a sequence of eight coin tosses would have 5 tails and 3 heads in any order.

(1) Coin toss events are independent of each other. The probability will be calculated as follows $0.4 \cdot 0.6 \cdot 0.6 \cdot 0.4 \cdot 0.4 \cdot 0.4 \cdot 0.6 \cdot 0.4 = 0.6^3 0.4^5 = 0.00221$. By symmetry, this is the same as the probability of getting the sequence $TTTTTHHH$ (or any other arrangements involving 3 heads and 5 tails) in exact order.

(2) The probability of getting 5 tails and 3 heads in any order can be calculated using the addition law for disjoint events:

$$P(3 \text{ heads}) = P(\{THHTTTHT\}) + P(\{THTHTTTH\})$$
$$+ P(\{TTHTHHTT\}) + \cdots$$
$$+ P(\{\text{other arrangements of } 3H \text{ and } 5T\})$$

In this case, we need to find the number of ways to permute 5 tails and 3 heads which is given by $\binom{8}{3} = 56$. Using our answer to (1), the probability is therefore:

$$P(3 \text{ heads}) = 56 \cdot 0.00221 = 0.12386$$

The total probability of binomial PMF is indeed one:

$$\sum_{k=0}^{n} p_X(k) = \sum_{k=0}^{n} \binom{n}{k} q^k (1-q)^{n-k} = (q + (1-q))^n = 1^n = 1$$

EXAMPLE 33.

Using the same unfair coin from the previous example, define a score to be (number of Heads) / (number of tosses). Find the probability that the score is between 0.4 and 0.8 when $N = 9$ coins are tossed. What is the answer if N is very large?

To get the score between 0.4 and 0.8, we need to get 4 to 7 heads. So the probability is

$$P(\text{score between } 0.4 \text{ and } 0.8) = P(4 \text{ or } 5 \text{ or } 6 \text{ or } 7 \text{ heads})$$
$$= P(4 \text{ heads}) + P(5 \text{ heads})$$
$$+ P(6 \text{ heads}) + P(7 \text{ heads})$$
$$= \sum_{k=4}^{7} \binom{9}{k} 0.6^k 0.4^{9-k}$$
$$= 0.8301$$

If N is very large, the probability will reach 1.

EXAMPLE 34.

You roll 5 dice and what is the probability that you get 3 sixes?

Let X denote the number of sixes showing. There is $\frac{1}{6}$ chance to roll six and $\frac{5}{6}$ chance to roll other numbers. The probability of getting 3 sixes is

$$P(X = 3) = \binom{5}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 = 0.032$$

3.2.5   Poisson Random Variable

Poisson random variable $X \sim \text{Pois}(\lambda)$ can be used to find the probability of some number of events happening in some specified time period given that the event happens at some constant rate of λ and the events occur independently. Its PMF is given by

$$p_X(n) = P(X = n) = e^{-\lambda} \frac{\lambda^n}{n!}$$

where $n = 0, 1, 2, \dots$. It calculates the probability that n events will happen at the specified time period. The parameter $\lambda > 0$ can be thought as the mean number of event occurrences happening in the given time period or rate (units: occurrences per time). This PMF works because

$$\sum_{n=0}^{\infty} p_X(n) = \sum_{n=0}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1$$

The probability that n events will happen in time interval t given the event rate r (the parameter is $\lambda = rt$) is

$$P(X = n) = e^{-rt} \frac{(rt)^n}{n!}$$

The possible examples of this random variable is not limited to the number of typos in a book, the number of radioactive decays in some time interval, and the number of people who failed exams.

Compared to binomial random variables which count discrete occurrences among discrete trials, the Poisson random variables count discrete occurrences among a continuous domain.
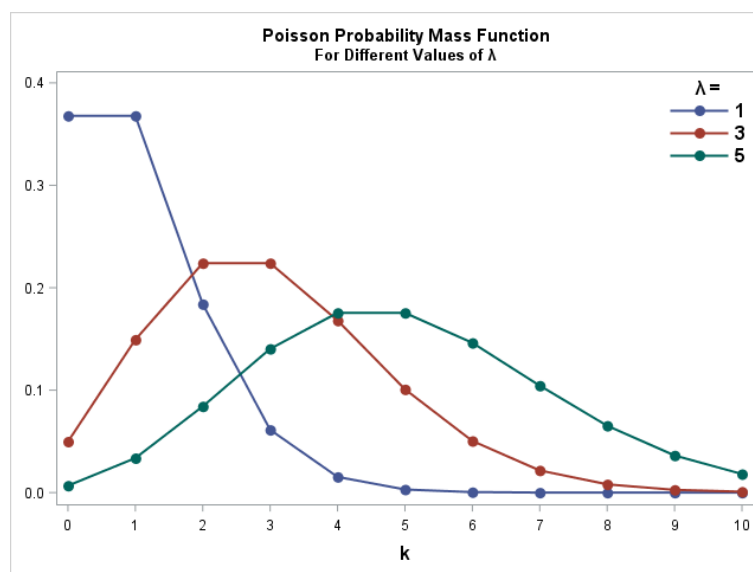

Figure 17: Poisson PMFs for different values of λ. The PMF with $\lambda = 3$ is peaked at $k = 3$ which makes sense because events with the mean number of occurrences $\lambda = 3$ per time frame of 10 will most likely happen at 3 times out of 10.

When the number of trials n of a binomial random variable approaches infinity and the probability of success p approaches zero, it can be approximated with Poisson random variable with parameter $\lambda = pn$. This is a very robust approximation.

Consider a binomial random variable $X \sim \text{Bin}(n, q)$, applying $\lambda = pn$ to its distribution gives

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

$$= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \tag{2}$$

$$= \frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-k} \tag{3}$$

Now as $n \to \infty$ and $p \to 0$, we have

$$\frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k} \approx 1 \quad \left(1 - \frac{\lambda}{n}\right)^{n} \approx e^{-\lambda} \quad \left(1 - \frac{\lambda}{n}\right)^{-k} \approx 1$$

This gives

$$P(X = k) \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

We get the Poisson approximation of X as expected. We can see that Poisson random variables are great for modelling occurrences of events that could happen a very large number of times but happen rarely. Sometimes Poisson is preferred over binomial as they make calculations easier.

EXAMPLE 35.

Birthday problem revisited: In a room of n people, approximate the probability that at least 2 people share the same birthday.

We will take a slightly different route from the last time. Imagine going over every possible pair of n people and marking it off as a success when both of them share the same birthday. The probability of success happening is $\frac{1}{365}$. There are $\binom{n}{2}$ possible pairs.

We can employ the Poisson approximation here. Let $X \sim \text{Pois}(\lambda)$ denote the number of pairs who share the same birthday where the average number of successful pairs is $\lambda = \binom{n}{2}\frac{1}{365}$. The probability that at least one pair share the same birthday is $P(X > 0) = 1 - P(X = 0)$ where $P(X = 0)$ is the probability that none of the pairs share the same birthday. We get

$$P(X > 0) = 1 - e^{-\lambda} = 1 - \exp\left(-\binom{n}{2}\frac{1}{365}\right)$$

As an example, suppose $n = 23$, $P(X > 0) = 0.5$.

The Poisson approximation can be generalized to more than two people. If we want to find the probability that at least three people share the same birthday, we only need modify the Poisson parameter which will be $\lambda = \binom{n}{3}\frac{1}{365^2}$ since we now need to choose three people and there is $\frac{1}{365^2}$ chance that three of them share the same birthday.

As a side note, binomial distribution cannot be used in this problem because it only works when events are independent which is not the case here. The probability of a person having a different birthday dependent on another person's birthday. But it is still "possible" to apply binomial distribution here since the Poisson's approximation works. For example if $n = 23$, the probability is

$$P(X > 0) = 1 - \left(\binom{n}{0} (p)^0 (1 - p)^n \right) = 0.50048$$

where $n = \binom{23}{2}$ and $p = \frac{1}{365}$. It is close enough to Poisson's approximation. This is only because n is large and p is small.

EXAMPLE 36.

You repeat an experiment 1000 times, each with an independent success rate of $\frac{1}{500}$. Find the probability of getting at most two successes.

Denote $n = 1000$ and $p = \frac{1}{500}$. One can use a binomial random variable to find probability:

$$P(\text{at most 2 success}) = P(0 \text{ success}) + P(1 \text{ success}) + P(2 \text{ success})$$

$$= \sum_{k=0}^{2} \binom{n}{k} p^k (1 - p)^{n-k}$$

$$= 0.67668$$

An even better way to solve is to approximate using a Poisson random variable $X \sim \text{Pois}(\lambda)$ where $\lambda = pn = 2$. This gives

$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

$$= e^{-\lambda} + e^{-\lambda}\frac{\lambda}{1} + e^{-\lambda}\frac{\lambda^2}{2} = 5e^{-2}$$

$$= 0.67668$$

This approximation works well because n is large and p is small.

3.2.6   Expectation Value

The expectation value of a discrete random variable X with corresponding PMF $p_X(x)$ is

$$E[X] = \sum_i x_i p_X(x_i)$$

It is the weighted average of all possible values X can take x_i. It is denoted by either $E[X]$ (or simply EX) or $\langle X \rangle$. It is different from the average which is the mean result obtained after repeating an experiment for some number of times. But the average will become the expectation value if the experiment is repeated for infinite number of times (law of large numbers).

EXAMPLE 37.

Let X represent the outcome of a dice roll. Find its expectation value.

The possible values of X are $x_i = 1, 2, \ldots, 6$ and the probability of rolling one of these values is $p_X(x_i) = \frac{1}{6}$ so the expectation value is:

$$E[X] = \sum_{i=1}^{6} x_i p_X(x_i) = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5$$

It represents the average result you get if you roll the dice for infinite number of times.

The expectation value for a Bernoulli random variable can be calculated in a straightforward manner as follows:

$$E[X \sim \text{Bern}(p)] = 1p + 0(1-p) = p$$

For geometric random variables, consider

$$f(x) = \sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad f'(x) = \sum_{n=0}^{\infty} nx^{n-1} = \left(\frac{1}{1-x}\right)^2$$

Let $x = 1 - q$,

$$f'(1-q) = \sum_{n=0}^{\infty} n(1-q)^{n-1} = \frac{1}{q^2}$$

multiplying both sides by q gives

$$q\sum_{n=0}^{\infty} n(1-q)^{n-1} = \sum_{n=0}^{\infty} nq(1-q)^{n-1} = \frac{1}{q}$$

This gives the closed formula for the expectation value of geometric random variables (remember that they start from $n = 1$, not $n = 0$):

$$E[X \sim \text{Geom}(q)] = \sum_{n=1}^{\infty} nq(1-q)^{n-1} = \frac{1}{q}$$

For binomial random variables, consider

$$\sum_{k=0}^{n} x^k p(k) = \sum_{k=0}^{n} \binom{n}{k}(qx)^k(1-q)^{n-k} = (1 - q + qx)^n$$

taking a derivative with respect to x gives

$$\sum_{k=0}^{n} kx^{k-1} p(k) = n(1 - q + qx)^{n-1}q$$

setting $x = 1$,

$$\sum_{k=0}^{n} kp(k) = n(1)^{n-1}q = nq$$

This gives the expectation value for a binomial random variable

$$E[X \sim \text{Bin}(n, q)] = nq$$

Finally, for Poisson random variables, we start with

$$f(x) = \sum_n e^{-\lambda} \frac{\lambda^n}{n!} x^n$$

applying the Taylor expansion for e^λ,

$$f(x) = e^{-\lambda} e^{\lambda x} = e^{\lambda x - \lambda}$$

the derivative of $f(x)$ is

$$f'(x) = \lambda e^{\lambda x - \lambda}$$

Notice that $f(1) = 1$, which is the total probability of Poisson PMF, and $f'(1) = \lambda$. This gives a Poisson random variable expectation

$$E[X \sim \text{Pois}(\lambda)] = \lambda$$

We see that the general strategy for finding closed-form expression for expectation is to start with a generating function of the following form

$$f(x) = \sum_n x^n p(n)$$

where x takes values in $\mathbb{N}$. $f(1) = 1$ represents the sum of probabilities. The derivative of this function is
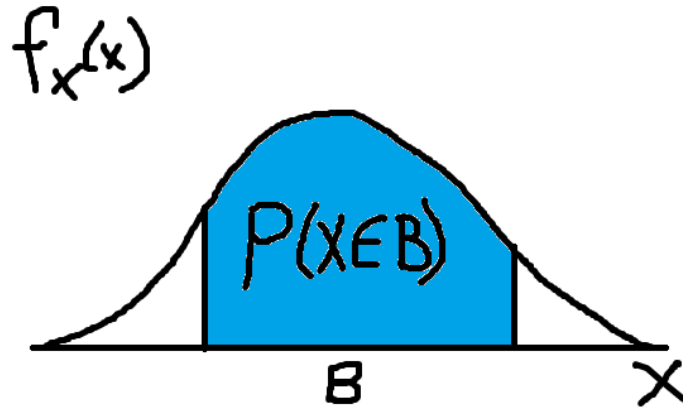
$$f'(x) = \sum_n n x^{n-1} p(n)$$

$f'(1)$ usually gives the expectation value.

3.3 Continuous Random Variables

A random variable X is said to be continuous if there exists a function $f_X(x)$ defined for all real numbers $x \in \mathbb{R}$ such that

$$P(X \in B) = \int_B f_X(x) dx$$

where B is a set of real numbers.

Figure 18: Illustration of $P(X \in B)$

The function $f_X(x)$ is known as a probability density function PDF of X. It can be obtained as a derivative of CDF:

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

given that $F_X(x)$ is continuous and differentiable everywhere. The requirements for $f_X(x)$ include

  i $f_X(x) \geq 0$ for all x

  ii $f_X(x)$ must be (piecewise) continuous for all x

  iii $\int_{-\infty}^{\infty} f_X(x)dx = 1$

$f_X(x)$ can have $f_X(x) > 1$ for some x values since it is not a probability function. The CDF of X can be readily obtained by

$$F_X(a) = P(X \leq a) = \int_{-\infty}^{a} f_X(x)dx$$

The probability of a continuous random variablele X in range (a, b) is

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = \int_{a}^{b} f_X(x)dx = F_X(b) - F_X(a)$$

It follows that the probability of X takes any specific a value is

$$P(X = a) = \int_{a}^{a} f_X(x)dx = 0$$

Note that even though the probability is zero, it is still possible for X to take a. This is because X can take infinitely many values in range $(-\infty, \infty)$ which makes the size of sample space infinite. The probability of landing on a specific a in this infinite sample space is very very low so the probability is zero but it does not make it impossible.

We will look over some special continuous random variables.

3.3.1 Continuous Uniform Random Variable

A continuous uniform distribution random variable $X \sim \text{Unif}(a, b)$ over the interval $[a, b]$ is described by the following PDF

$$f_X(x) = \begin{cases} c & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

where c is some normalization constant. We can calculate it by follows:

$$\int_a^b c\, dx = c(b - a) = 1$$

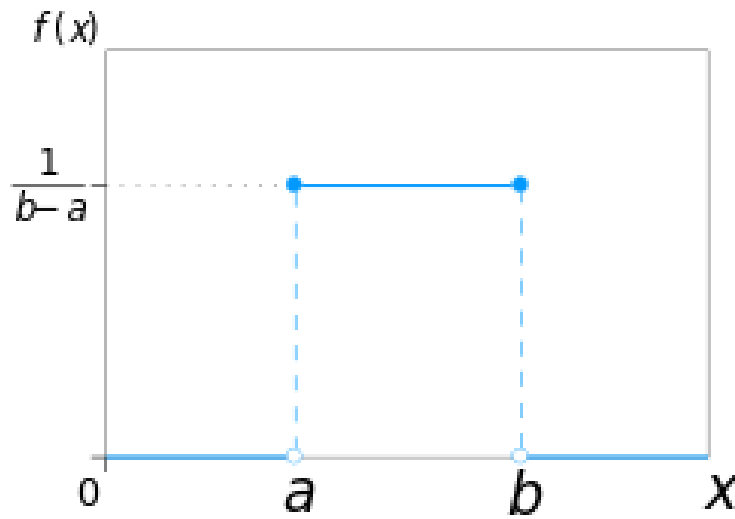which gives $c = \frac{1}{b-a}$.

Figure 19: PDF for Continuous Uniform Distribution

It can be used to model outcomes that lie in some boundary. The total probability is 1 as expected because the area of the rectangular region is

$$(b - a)\frac{1}{b - a} = 1$$

Its CDF can be calculated by

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(x) dx = \int_a^x \frac{1}{b - a} dx = \frac{x - a}{b - a}$$

for $x \in [a, b]$. The full CDF for this distribution is:

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

3.3.2 Exponential Random Variable

An exponential random variable $X \sim \text{Exp}(\lambda)$ of parameter λ is described by the PDF

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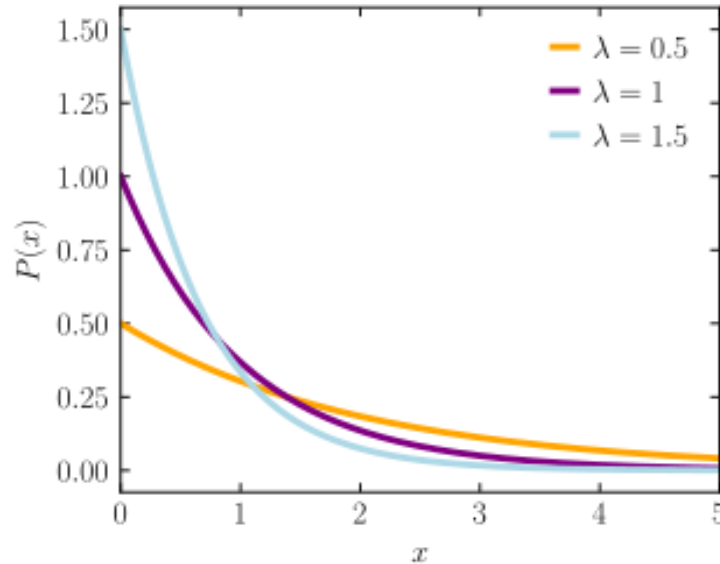

Figure 20: PDF for Exponential Distribution

This distribution is often used to find waiting times for periodic events. As an example suppose in a certain country (Japan I think), there is an earthquake that happens every ten years. Define λ (in fact it is the same one used in Poisson's) as the number of earthquakes happening every year which is $\lambda = \frac{1}{10}$. Let the random variable X denote waiting time in years. The probability that the earthquake happens in less than five years is

$$P(X \leq 5) = \int_0^5 \lambda e^{-\lambda x} dx = 0.393$$

and the probability that you need to wait more than a year to see the earthquake happen is

$$P(X > 1) = 1 - P(X \leq 1) = 1 - \int_0^1 \lambda e^{-\lambda x} dx = 0.9048$$

Note that you can calculate the same probability using a Poisson random variable with the same parameter. Redefine X as the number of earthquakes happening in a one-year timespan. The probability that there are no earthquakes in one year is

$$P(X = 0) = e^{-\lambda} = 0.9048$$

It gives the same result! An exponential distribution's "waiting time" does not have any events happening, hence in the Poisson distribution we set the number of events happening in that time period to zero $X = 0$.

The other applications of exponential distributions include finding the half-life of radioactive material. These problems can be modelled as finding time a such that

$$P(X \geq a) = \frac{1}{2}$$

a can be found by

$$e^{-\lambda a} = \frac{1}{2}$$

which gives $a = \frac{\ln(2)}{\lambda}$.

The total probability of the exponential distribution is indeed one:

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_{0}^{\infty} \lambda e^{-\lambda x}dx = -e^{-\lambda x}\big|_{0}^{\infty} = 1$$

Exponential distributions also have memoryless property like geometric random variables. The waiting time for a future event does not depend on elapsed time:

$$P(X \geq s + t \mid X \geq s) = P(X \geq t)$$

Proof. Evaluating $P(X \geq t)$ gives

$$P(X \geq t) = \int_{t}^{\infty} \lambda e^{-\lambda x}dx = -e^{-\lambda x}\big|_{t}^{\infty} = e^{-\lambda t}$$

It also follows that

$$P(X \geq s) = e^{-\lambda s} \quad P(X \geq s + t) = e^{-\lambda(s+t)}$$

Using the definition of conditional probability to calculate $P(X \geq s + t \mid X \geq s)$ gives

$$P(X \geq s + t \mid X \geq s) = \frac{P(X \geq s + t)}{P(X \geq s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}$$

this is the same as $P(X \geq t)$ which concludes the proof. $\square$

As a supplement, we will show how Poisson distribution is used to derive exponential distribution. Let X denote number of years. We saw earlier in the earthquake example that the probability that no earthquakes happen in one year is $e^{-\lambda}$. Now the probability that no earthquakes happen in x years is just

$$P(X > x) = \underbrace{e^{-\lambda}e^{-\lambda}\cdots e^{-\lambda}}_{x \text{ times}} = e^{-\lambda x}$$

This allows us to define exponential distribution's CDF which is given by

$$F_X(x) = P(X \leq x) = 1 - P(X > x) = 1 - e^{-\lambda x}$$

Finally we obtain exponential PDF by differentiating the CDF:

$$f_X(x) = \frac{d}{dx}F_X(x) = \lambda e^{-\lambda x}$$

3.3.3   Normal Random Variable

A normal random variable $X \sim \text{Norm}(\mu, \sigma^2)$ of mean μ and variance σ^2 is described by the following PDF:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$
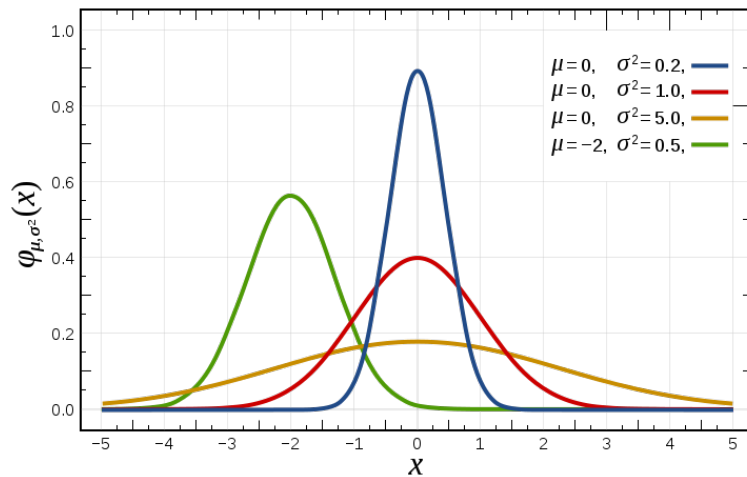
It produces the famous bell curve

Figure 21: Normal PDF. μ specifies the center of mass of the distribution and σ^2 controls how wide the curve is. Bigger the σ^2 is, more flatter the curve becomes.
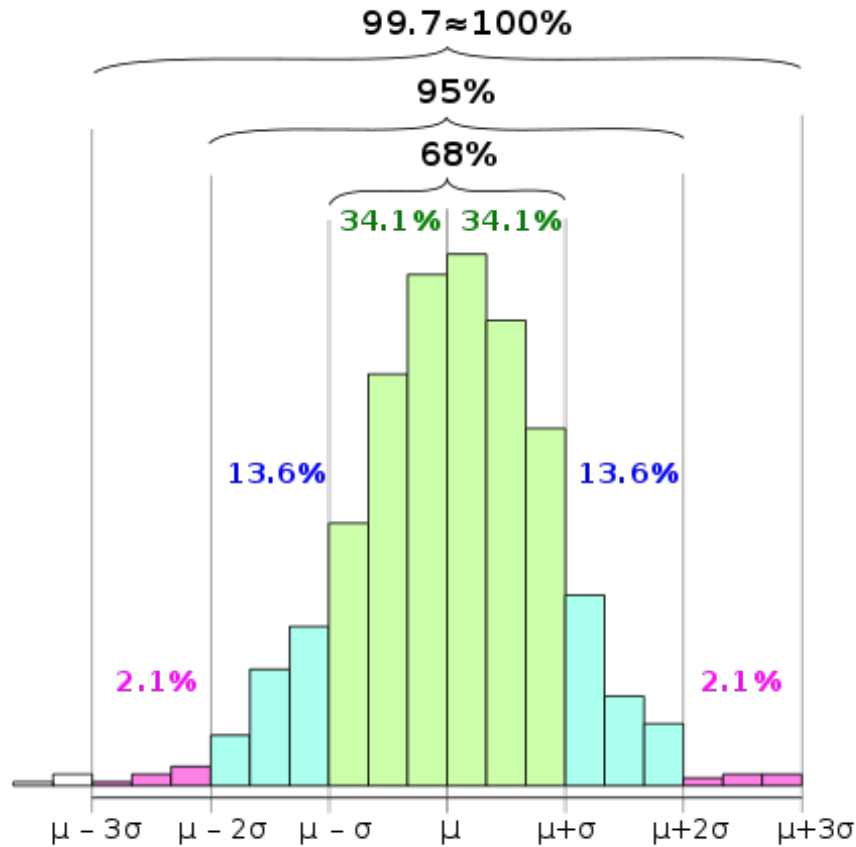
Figure 22: Distribution of mass in Gaussian. One standard derivation σ left and right from mean value μ will give about 68% of the overall mass. Two standard derivations left and right from the mean will give about 95% of the overall mass.

Like other distributions, the total probability of this distribution is indeed one.

Proof.

$$\int_{-\infty}^{\infty} f_X(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$$

Change of variables using $t = x - \mu$ gives

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt$$

Denote the integral $\int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt$ by I. Integrating I^2 gives

$$I^2 = \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2\sigma^2}\right) dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2} - \frac{y^2}{2\sigma^2}\right) dxdy$$

We switch to polar coordinates

$$= \int_0^\infty \int_0^{2\pi} \exp\left(-\frac{r^2}{2\sigma^2}\right) r d\theta dr$$

$$= 2\pi \int_0^\infty \exp\left(-\frac{r^2}{2\sigma^2}\right) r dr$$

We make another change of variables using $u = \frac{r^2}{2\sigma^2}$ which gives

$$= 2\pi \int_0^\infty e^{-u} \sigma^2 du$$

$$= 2\pi\sigma^2$$

This gives $I = \sigma\sqrt{2\pi}$. Now

$$\int_{-\infty}^\infty f_X(x)dx = \frac{1}{\sigma\sqrt{2\pi}}I = \frac{1}{\sigma\sqrt{2\pi}}\sigma\sqrt{2\pi} = 1$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

The CDF for normal distribution is

$$F_X(a) = P(X \le a) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^a \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$$

There is no closed-form expression for $F_X(a)$. However, we can define standard normal Norm$(0,1)$ whose PDF is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

and its CDF is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{x^2}{2}\right) dx$$

Standard normal's CDF $\phi(x)^2$ can be used to calculate $F_X(a)$:
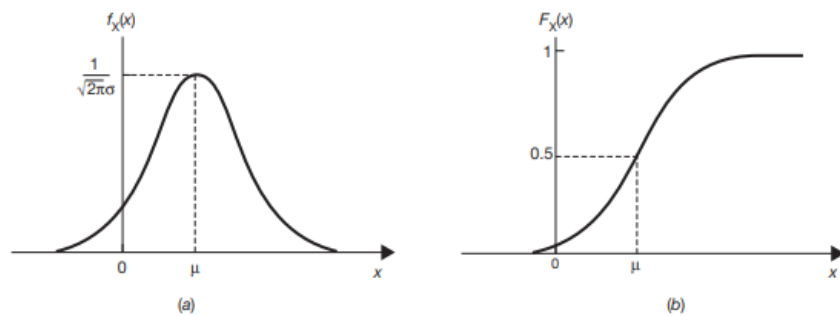
$$F_X(a) = \phi\left(\frac{a-\mu}{\sigma}\right)$$

Figure 23: Standard Normal's PDF and CDF

---

[2]In terms of error round function erf, ϕ can be written as $\phi(x) = \frac{1+\text{erf}(x/\sqrt{2})}{2}$.

Normal random variable $X \sim \text{Norm}(\mu, \sigma^2)$ can be scaled to standard normal $Z \sim \text{Norm}(0, 1)$ with $Z = \frac{X-\mu}{\sigma}$.

Proof. The probability of $Z \leq t$ is

$$P(Z \leq t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-\frac{z^2}{2}\right) dz \quad \star$$

Since $X = \sigma Z + \mu$, $Z \leq t$ is same as $X \leq \sigma t + \mu$. So the probability of $X \leq \sigma t + \mu$ is

$$P(X \leq \sigma t + \mu) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\sigma t + \mu} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$$

Change of variables using $z = \frac{x-\mu}{\sigma}$[3] will convert this integral to the one marked with $\star$. $\qquad \square$

Normal distributions have a broad range of applications everywhere. It is often used for random variables with unknown distribution patterns.

EXAMPLE 38.

A group of programmers is trying to fix a buggy random number generator. They found that the mean of all generated random numbers is 310 with standard deviation is 23. What is the probability that it will generate a number somewhere between 250 and 300?

Let the normal random variable $X \sim \text{Norm}(310, 23^2)$ denote the random number generated. We want to calculate $P(250 \leq X \leq 300)$. We can simplify the calculation using this substitution $Z = \frac{X-\mu}{\sigma}$ so the probability can be looked up using a standard normal table.

$$P(250 \leq X \leq 300) = P(\frac{250 - 310}{23} \leq \frac{X - \mu}{\sigma} \leq \frac{300 - 310}{23}) = P(-2.61 \leq Z \leq -0.43)$$

Since the normal distribution is symmetric, $P(-2.61 \leq Z \leq -0.43)$ is same as $P(0.43 \leq Z \leq 2.61)$. The probability is

$$P(0.43 \leq Z \leq 2.61) = P(0 \leq Z \leq 2.61) - P(0 \leq Z \leq 0.43) \approx 0.4955 - 0.1664 = 0.3291$$

EXAMPLE 39.

A prisoner plays with a laser pointer in his jail cell. He randomly directs the laser beam at the wall one meter away from him.

---

[3]In statistics, z is known as "z-score." It measures how far x deviates from mean μ in units of σ.
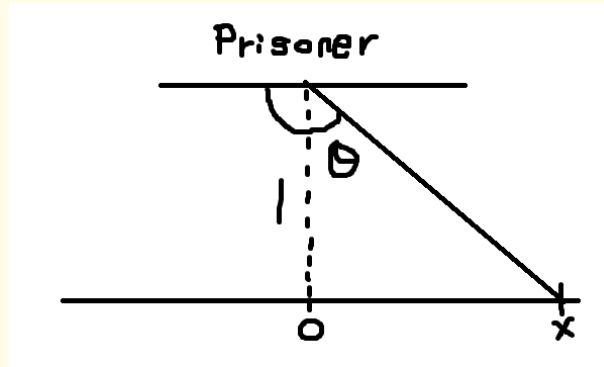
Figure 24: Prisoner Playing with Laser

The laser beam angle θ is uniform in $[0, \pi]$. A security guard wants to find the most likely distance x from origin O where the laser beam lands.

Let random variable X denote the distance from the origin. We need to find CDF for X:

$$F_X(x) = P(X \leq x)$$

We can calculate $P(X \leq x)$ using geometry. x is some arbitrary distance from origin. From the diagram when the laser beam hits x, its angle is $\arctan x + \frac{\pi}{2}$. When $X \leq x$, it follows that the angle $\theta \leq \arctan x + \frac{\pi}{2}$. The probability that $X \leq x$ is:

$$P(X \leq x) = P(\theta \leq \arctan x + \frac{\pi}{2}) = \frac{1}{\pi}\left(\arctan x + \frac{\pi}{2}\right) = \frac{1}{\pi}\arctan x + \frac{1}{2}$$

The PDF for X can be obtained by differentiating $F_X(x)$:

$$f(x) = \frac{d}{dx}F_X(x) = \frac{1}{\pi(1 + x^2)}$$
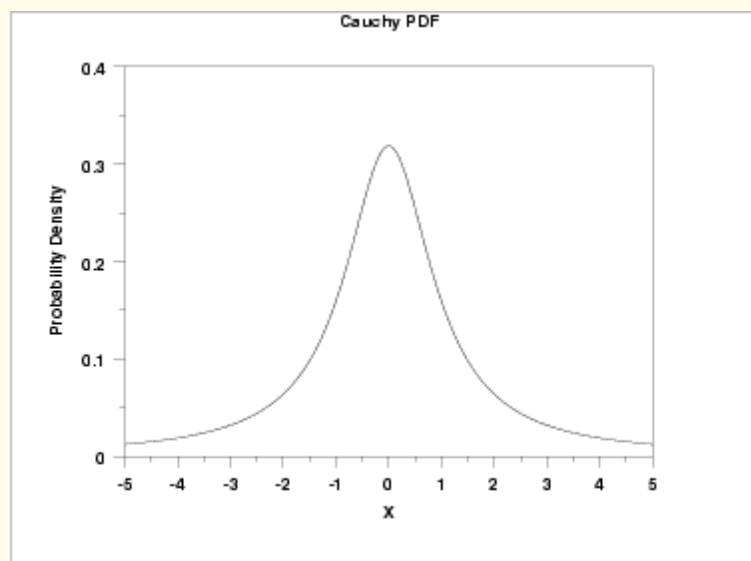
This is also the PDF for the standard Cauchy distribution.

Figure 25: Standard Cauchy PDF

The laser beam will most likely hit somewhere around $x = 0$.

3.3.4   Expectation Value

For continuous random variables X with PDF $f(x)$, its expectation is defined as

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

Expectation for continuous uniform random variables can be calculated as follows:

$$E[X \sim \text{Unif}(a,b)] = \int_{-\infty}^{\infty} x f(x) dx = \int_{a}^{b} \frac{x}{b-a} dx = \frac{a+b}{2}$$

For exponential random variables, straightforward integration by parts gives

$$
\begin{aligned}
E[X \sim \text{Exp}(\lambda)] &= \int_{0}^{\infty} x \lambda e^{-\lambda x} dx \\
&= -x e^{-\lambda x} \big|_{0}^{\infty} + \int_{0}^{\infty} e^{-\lambda x} dx \\
&= \frac{1}{\lambda}
\end{aligned}
$$

For normal random variables,

$$E[X \sim \text{Norm}(\mu, \sigma^2)] = \mu$$

I leave the calculation details for the readers.

3.4   More on Expectation

Theorem. *If a continuous random variable $X \geq 0$, then*

$$EX = \int_{0}^{\infty} P(X \geq t) dt$$

Proof.

$$EX = \int_{0}^{\infty} P(X \geq t) dt = \int_{0}^{\infty} 1 - F_X(t) dt$$

where $F_X(t)$ is CDF of X. Integrating by parts using $u = 1 - F_X(t)$ and $dv = dt$ gives

$$EX = t(1 - F_X(t)) \big|_{0}^{\infty} + \int_{0}^{\infty} t f_X(t) dt$$

where $f_X(t) = \frac{d}{dt} F_X(t)$ is PDF for X. Provided that $\lim_{t \to \infty} t(1 - F_X(t)) = 0$, we are left with

$$EX = \int_{0}^{\infty} t f_X(t) dt$$

$\square$

If X is a discrete random variable that takes values $0, 1, 2, \ldots$,

$$EX = \sum_{a=0}^{\infty} P(X > a)$$

Theorem. *Law of the Unconscious Statistician: Let $g : \mathbb{R} \to \mathbb{R}$ be an only function and X is a continuous random variable, then $g(X)$ is a new random variable. It follows that the expectation value for $g(X)$ is*

$$Eg(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

If X is discrete random variable,

$$Eg(X) = \sum_{i} g(x_i)p(x_i)$$

Proof. If $g \geq 0$ then $g(X) \geq 0$. The expectation for $g(X)$ is

$$Eg(X) = \int_{0}^{\infty} P(g(x) \geq t)dt$$

Define a set $B_t = \{x : g(x) \geq t\}$,

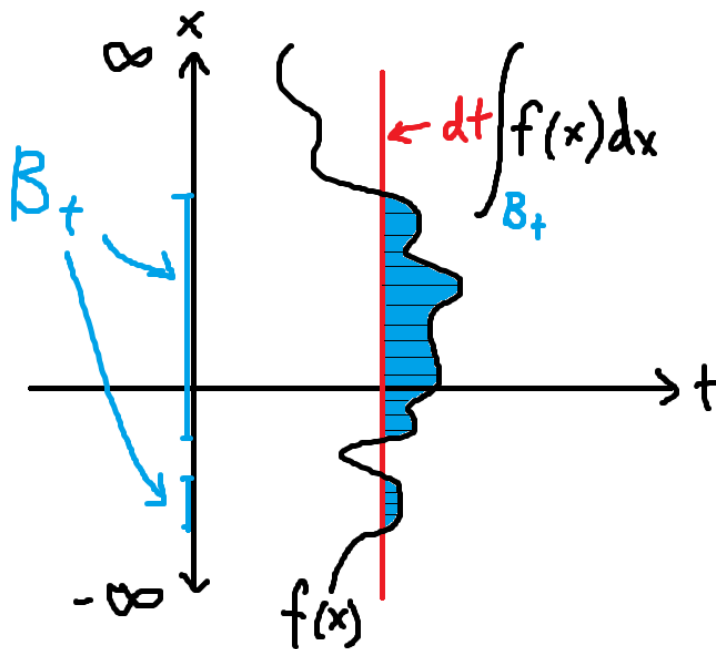$$Eg(X) = \int_{0}^{\infty} \int_{B_t} f(x)dx \, dt$$

Figure 26: Visualization of the integral $\int_{0}^{\infty} \int_{B_t} f(x)dx \, dt$

Changing the integration order gives

$$Eg(X) = \int_{-\infty}^{\infty} \int_{0}^{g(x)} f(x)dt \, dx = \int_{-\infty}^{\infty} g(x)f(x)dx$$

$\square$

There is an important special case to consider: Let a and b denote constants. Suppose $g(X) = aX + b$,

$$E(aX + b) = \int (ax + b)f(x)dx = a \int xf(x)dx + b \int f(x)dx = aEX + b$$

(The whole integral of the PDF is one: $\int_{-\infty}^{\infty} f(x)dx = 1$.) This shows that expectation is linear.

EXAMPLE 40.

A particle of mass 1g has a random velocity X that is uniformly distributed between 3cm/s and 8cm/s. Find the probability density of the particle's kinetic energy $T = \frac{1}{2}X^2$. Find the mean of T in two ways: using the PDF for T and the Law of the Unconscious Statistician.

X is a continuous uniform random variable with the following CDF:

$$F_X(x) = P(X \le x) = \begin{cases} 0 & x < 3 \\ \frac{x-3}{5} & 3 \le x \le 8 \\ 1 & x > 8 \end{cases}$$

The CDF for T is:

$$F_T(t) = P(T \le t) = P\left(\frac{1}{2}X^2 \le t\right) = P(X \le \sqrt{2t}) = \begin{cases} 0 & t < \frac{9}{2} \\ \frac{\sqrt{2t}-3}{5} & \frac{9}{2} \le t \le 32 \\ 1 & t > 32 \end{cases}$$

The PDF for T is:

$$f_T(t) = \frac{d}{dt}F_T(t) = \begin{cases} 0 & t < \frac{9}{2} \\ \frac{1}{5\sqrt{2t}} & \frac{9}{2} \le t \le 32 \\ 1 & t > 32 \end{cases}$$

The mean for T is using the PDF directly is:

$$ET = \int_{\frac{9}{2}}^{32} t \frac{1}{5\sqrt{2t}} dt = \frac{97}{6}$$

or using the Law of the Unconscious Statistician:

$$ET = E[\frac{1}{2}X^2] = \int_{3}^{8} \left(\frac{1}{2}x^2\right)\frac{1}{5}dx = \frac{97}{6}$$

They both give the same answer!

EXAMPLE 41.

Stanislaw is collecting coupons. Each day he receives randomly one of n distinct coupons with equal probabilities (independently of other days). Let T be the number of days it takes Stanislaw to obtain a complete set. Compute the expected value of T.

Let T_i be the number of days until Stanislaw has i different coupons. We let $T_0 = 0$, and clearly $T_1 = 1$. We are interested in $T = T_n$. Let $X_i = T_i - T_{i-1}$, so $T = X_1 + X_2 + \cdots + X_n$. Once Stanislaw has $i - 1$ coupons, each day there is probability $\frac{n-i+1}{n}$ of getting a new coupon. These attempts to get a new coupon are independent until the next success, so $X_i \sim \text{Geom}(q)$ with parameter $q = \frac{n-i+1}{n}$. The expectation for T is

$$E[T] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} \approx n \log n$$

3.5   Moment and Variance

The nth moment of a random variable X is defined as

$$E(X^n) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x^n f(x) dx & X \text{ continuous} \\[2em] \displaystyle\sum_i x_i^n p(x_i) & X \text{ discrete} \end{cases}$$

Note that $E(X^n) = (EX)^n$ is not true all the time. From the definition, it follows that mean EX is the first moment of X. EX is usually denoted by μ.

The variance (or second central moment) of X is defined as

$$\text{Var}(X) = \sigma^2 = E((x - \mu)^2)$$

$\text{Var}(X)$ is always greater than or equal to zero. The standard deviation is just the square root of variance: $\sigma = \sqrt{\text{Var}(X)}$. They both measure the spread of distribution from the mean.

Note that the following is equivalent

$$\text{Var}(X) = E(X^2) - (EX)^2$$

Proof.

$$\begin{aligned} \text{Var}(X) &= E((x - \mu)^2) \\ &= E(x^2 - 2x\mu + \mu^2) \\ &= \int (x^2 - 2x\mu + \mu^2) f(x) dx \\ &= \int x^2 f(x) dx - 2\mu \int x f(x) dx + \mu^2 \int f(x) dx \\ &= E(X^2) - 2\mu EX + \mu^2 \\ &= E(X^2) - 2(EX)^2 + (EX)^2 \\ &= E(X^2) - (EX)^2 \end{aligned}$$

$\square$

Variances for some random variables are given in the below table

Random variable X	$\text{Var}(X)$
$\text{Bin}(n, p)$	$np(1-p)$
$\text{Pois}(\lambda)$	λ
$\text{Exp}(\lambda)$	$\frac{1}{\lambda^2}$
$\text{Norm}(\mu, \sigma^2)$	σ^2

Theorem. *Let a and b denote constants,*

$$\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$$

or variance is independent of b added to X (ie. invariant of location) and if X is scaled by a, variance is scaled by a^2.

Proof. First, we prove that $\operatorname{Var}(X + b) = \operatorname{Var}(X)$,

$$\operatorname{Var}(X + b) = E((X + b)^2) - (E(X + b))^2$$

Since expectation is linear operator: $E(X + b) = EX + b$,

$$
\begin{aligned}
&= E(X^2 + 2bX + b^2) - (EX + b)^2 \\
&= E(X^2) + 2b(EX) + b^2 - ((EX)^2 + 2b(EX) + b^2) \\
&= E(X^2) + 2b(EX) + b^2 - (EX)^2 - 2b(EX) - b^2 \\
&= E(X^2) - (EX)^2 \\
&= \operatorname{Var}(X)
\end{aligned}
$$

Next, we prove that $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$,

$$
\begin{aligned}
\operatorname{Var}(aX) &= E((aX)^2) - (E(aX))^2 \\
&= a^2 E(X^2) - a^2 (EX)^2 \\
&= a^2 (E(X^2) - (EX)^2) \\
&= \operatorname{Var}(X)
\end{aligned}
$$

$\square$

4   Joint Distributions

4.1   Discrete Case

Let X, Y be discrete random variables, the function

$$p_{XY}(x, y) = P(X = x, Y = y)$$

is called the joint PMF of X and Y.

Joint PMFs share many properties with ordinary PMFs. For example, the total probability is one:

$$\sum_i \sum_j p_{XY}(x_i, y_i) = 1$$

To calculate the probability that X takes values in range $[x_1, x_2]$ and Y takes values in range $[y_1, y_2]$:

$$P((x_1 \le X \le x_2) \cap (y_1 \le Y \le y_2)) = \sum_{x=x_1}^{x_2} \sum_{y=y_1}^{y_2} p_{XY}(x, y)$$

We can obtain PMF for X alone from the joint PMF. It is called the marginal PMF of X and is defined as

$$p_X(x) = P(X = x) = \sum_j p_{XY}(x, y_j)$$

Similarly for Y,

$$p_Y(y) = P(Y = y) = \sum_i p_{XY}(x_i, y)$$

The below shows an example joint PMF:

Y \ X	1	2	3	
1	1/4	1/8	1/8	1/2
4	1/6	1/3	0	1/2
	10/24	11/24	3/24	p_X \ p_Y

All of its entries sum up to one:

$$\frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{6} + \frac{1}{3} + 0 = 1$$

Also, notice the total for p_X and p_Y add up to one as well.

EXAMPLE 42.

You toss 4 dice. Let random variables X denote the number of distinct results and Y denote the minimal result. Find $P(X = 1, Y = 3)$ and $P(X = 4, Y = 3)$.

$$P(1,3) = P(\{3,3,3,3,3\}) = \frac{1}{6^4}$$

$$P(4,3) = P(\text{roll } 3,4,5,6 \text{ in some order}) = \frac{4!}{6^4}$$

4.2 Continuous Case

Let X, Y be continuous random variables, the function $f_{XY}(x,y)$ is called the joint PDF of X and Y. The probability can be calculated as follows

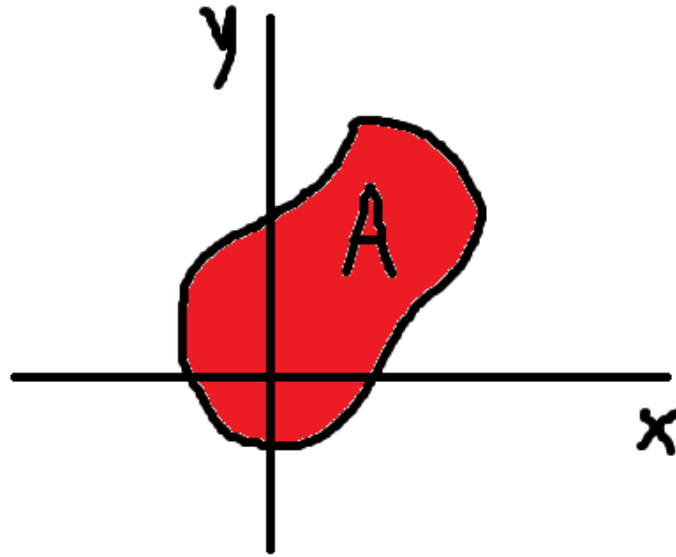$$P((X,Y) \in A) = \iint_A f_{XY}(x,y) dx\, dy$$

Figure 27: Illustration of A

Joint PDFs share many properties with ordinary PDFs. For example, $f_{XY}(x,y) \geq 0$ everywhere and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) dx\, dy = 1$$

To find the probability that X takes values in range $[x_1, x_2]$ and Y takes values in range $[y_1, y_2]$:

$$P((x_1 \leq X \leq x_2) \cap (y_1 \leq Y \leq y_2)) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{XY}(x,y) dx\, dy$$

Marginal PDF for X:
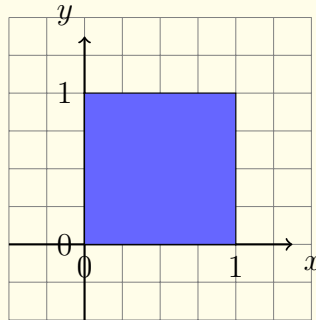
$$f_X(x) = \int_{-\infty}^{\infty} f(x,y) dy$$

Marginal PDF for Y:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y) dx$$

EXAMPLE 43.
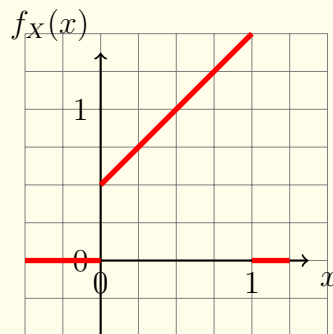
Consider the following joint PDF:

$$f_{XY}(x,y) = \begin{cases} x+y & 0 \le x,y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the marginal PDF for X.

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy = \int_0^1 x+y \ dy = x + \frac{1}{2}$$

4.3   Expectation

If X and Y are random variables and g is a function that takes these random variables, the expectation of $g(X,Y)$ is

$$Eg(X,Y) = \sum_i \sum_j g(x_i, y_i)p(x_i, y_i)$$

for discrete random variables or

$$Eg(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dx\ dy$$

for continuous random variables. For example, if $g(X,Y) = X + Y$, when X and Y are continuous,

$$E[X+Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y)f(x,y)dx\,dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x,y)dx\,dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x,y)dx\,dy$$

$$= EX + EY$$

It turns out that the linearity of expectation still holds for multiple random variables:

$$E\left(\sum_i a_i X_i\right) = \sum_i a_i E(X_i)$$

4.4   Independence

The random variables X and Y are independent if for any $a, b \in \mathbb{R}$,

$$P(X \le a, Y \le b) = P(X \le a)P(Y \le b)$$

ie. events $X \le a$ and $Y \le b$ are independent. In terms of PMF and PDF, X and Y are independent if

$$p_{XY}(x,y) = p_X(x)p_Y(y)$$

for discrete case and

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

for continuous case.

For any functions $g(X)$ and $h(Y)$, there exists a simple formula for calculating $E[g(X)h(Y)]$ when X and Y are independent.

$$E[g(X)h(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f(x,y)dx\,dy$$

Since X and Y are independent, $f(x,y) = f_X(x)f_Y(y)$,

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dx\,dy$$

$$= \left(\int_{-\infty}^{\infty} g(x)f_X(x)dx\right)\left(\int_{-\infty}^{\infty} h(y)f_Y(y)dy\right)$$

$$= E[g(X)]E[h(Y)]$$

In particular, $E[XY] = E[X]E[Y]$. This works for discrete random variables as well.

EXAMPLE 44.

Consider the following joint PMF

$$p_{XY}(x,y) = \frac{x^2 y}{30}$$

for discrete random variables $X = 1, 2$ and $Y = 1, 2, 3$. Show that X and Y are

---

independent.

The marginal PMF for X is

$$p_X(x) = \sum_{y=1}^{3} \frac{x^2 y}{30} = \frac{x^2}{5}$$

The marginal PMF for Y is

$$p_Y(y) = \sum_{x=1}^{2} \frac{x^2 y}{30} = \frac{y}{6}$$

Since

$$p_{XY}(x, y) = p_X(x)p_Y(y) = \frac{x^2 y}{30}$$

So X and Y are independent.

EXAMPLE 45.

The time T (in hours past noon) until the arrival of the first taxi has Exp(6) distribution, and the time B until first bus is independent with Exp(4) distribution. What is the joint probability density function of T and B? Find the probability that the first taxi arrives before the first bus. If you arrive at noon and take the first bus or taxi (whichever arrives first), what is the distribution of your waiting time?

The joint PDF for T and B is:

$$f_{TB}(t, b) = f_T(t)f_B(b) = \begin{cases} 6e^{-6t}4e^{-4t} & t, b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The probability that the first taxi arrives before the first bus $T < B$ is:

$$P(T < B) = \iint_{t<b} f_{TB}(t, b)db\, dt = \int_0^\infty \int_t^\infty 6e^{-6t}4e^{-4t}db\, dt = \frac{3}{5}$$

To find the distribution for waiting time, denote a random variable $X = \min(T, B)$.

$$P(X > t) = P(T > t, B > t)$$

Since T and B are independent,

$$= P(T > t)P(B > t)$$
$$= \left(1 - \int_0^t 6e^{-6t'}dt'\right)\left(1 - \int_0^t 4e^{-4t'}dt'\right)$$
$$= \left(1 - (1 - e^{-6t})\right)\left(1 - (1 - e^{-4t})\right)$$
$$= e^{-6t}e^{-4t}$$
$$= e^{-10t}$$

The CDF for X is therefore

$$P(X \leq t) = 1 - P(X > t) = 1 - e^{-10t}$$

It looks like the CDF for Exp(10).

EXAMPLE 46.

If X, Y are independent Norm$(0, 1)$ random variables, what is the distribution of $R = \sqrt{X^2 + Y^2}$?

The joint density is given by

$$f_{XY})(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right)$$

We first calculate CDF for R:

$$
\begin{aligned}
F_R(a) &= P(\sqrt{X^2 + Y^2} \leq a) \\
&= P(X^2 + Y^2 \leq a^2) \\
&= \iint_{x^2 + y^2 \leq a^2} \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right) dx \, dy
\end{aligned}
$$

Changing to polar coordinates gives

$$
\begin{aligned}
&= \frac{1}{2\pi} \int_0^{2\pi} \int_0^a \exp\left(-\frac{r^2}{2}\right) r dr \, d\theta \\
&= 1 - \exp\left(-\frac{a^2}{2}\right)
\end{aligned}
$$

PDF can be calculated via differentiating the CDF:

$$f_R(a) = \frac{d}{da}\left(1 - \exp\left(-\frac{a^2}{2}\right)\right) = a \exp\left(-\frac{a^2}{2}\right)$$

EXAMPLE 47.

A point (X, Y) is uniformly chosen in the unit disc $x^2 + y^2 \leq 1$. (a) Determine the marginal PDF of X and Y and justify whether X and Y are independent. (b) Find the PDF of its distance from the origin $R = \sqrt{X^2 + Y^2}$. Compute the expected distance from the origin.

(a) The PDF for points uniformly distributed on the unit disk is given by:

$$f_{XY}(x, y) = \begin{cases} c & x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The normalization factor c is calculated by integrating over region $x^2 + y^2 \leq 1$ (this is same as finding the area of the unit circle):

$$\iint_{x^2+y^2\leq1} c dx\, dy = c \int_0^{2\pi} \int_0^1 r dr\, d\theta = c\pi = 1$$

which gives $c = \frac{1}{\pi}$.

The marginal PDFs for X and Y are:

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{2}{\pi}\sqrt{1-x^2}$$

for $x \in [-1,1]$.

$$f_Y(y) = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{1}{\pi} dx = \frac{2}{\pi}\sqrt{1-y^2}$$

for $y \in [-1,1]$.

Since $f_X(x)f_Y(y) \neq \frac{1}{\pi}$, X and Y are not independent.

(b) The CDF for R is:

$$F_R(a) = P(X^2 + Y^2 \leq a^2) = \iint_{x^2+y^2\leq a^2} f_{XY}(x,y) dx\, dy = \int_0^{2\pi} \int_0^a \frac{1}{\pi} r dr\, d\theta = a^2$$

and its PDF is:

$$f_R(a) = \frac{d}{da}(a^2) = 2a$$

So the expected distance from origin is given by:

$$ER = \int_0^1 r f_R(r) dr = \int_0^1 2r^2 dr = \frac{2}{3}$$

EXAMPLE 48.

Let X, Y be uniform on the square $[0,2] \times [0,2]$. Find the distribution of the ratio $Z = X/Y$.

We are given that

$$f_{XY}(x,y) = \begin{cases} c & x \in [0,2] \text{ and } y \in [0,2] \\ 0 & \text{otherwise} \end{cases}$$

where c is calculated by:

$$\int_0^2 \int_0^2 c dx\, dy = 4c = 1$$

We first calculate the CDF for Z:

$$F_Z(a) = P(Z \le a) = P\left(\frac{X}{Y} \le a\right) = P(X \le aY)$$

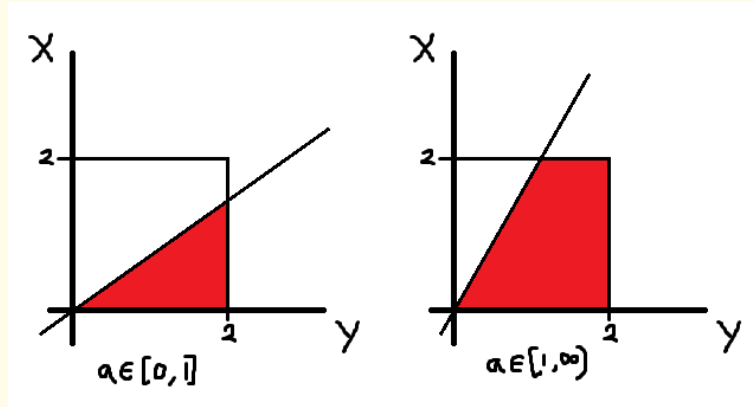To find $P(X \le aY)$, we need to consider two cases: $a \in [0, 1]$ and $a \in [1, \infty)$.

Figure 28: Plot of the line $X = aY$ and integration regions in two cases.

For $a \in [0, 1]$,

$$P(X \le aY) = \int_0^2 \int_0^{ay} \frac{1}{4} dx \, dy = \frac{a}{2}$$

For $a \in [1, \infty)$,

$$P(X \le aY) = 1 - \int_0^2 \int_0^{\frac{x}{a}} \frac{1}{4} dy \, dx = 1 - \frac{1}{2a}$$

4.5 Covariance and Correlation Coefficient

Covariance between two random variables X and Y is defined as

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)]$$

In particular, $\text{Cov}(X, X) = \text{Var}(X)$. There is another formula for covariance:

$$
\begin{aligned}
\text{Cov}(X, Y) &= E[(X - EX)(Y - EY)] \\
&= E[XY - (EX)Y - X(EY) + (EX)(EY)] \\
&= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\
&= E[XY] - E[X]E[Y]
\end{aligned}
$$

[4] It follows that if X and Y are independent, $\text{Cov}(X, Y) = 0$ since $E[XY] = E[X]E[Y]$.

EXAMPLE 49.

Consider discrete random variables X and Y, Their joint PMF is specified below

---

[4] $E[XY] = \sum_i \sum_j x_i y_j p_{XY}(x_i, y_j)$

X \ Y	0	1
−1	0	1/4
1	1/2	1/4

Find $\text{Cov}(X,Y)$.

We have

$$EX = 0 \cdot \frac{1}{2} + 1 \cdot \left(\frac{1}{4} + \frac{1}{4}\right) = \frac{1}{2}$$

$$EY = -1 \cdot \frac{1}{4} + 1 \cdot \left(\frac{1}{2} + \frac{1}{4}\right) = \frac{1}{2}$$

$$E[XY] = -1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = 0$$

Covariance of X and Y is therefore

$$\text{Cov}(X,Y) = E[XY] - EXEY = 0 - \frac{1}{2} \cdot \frac{1}{2} = -\frac{1}{4}$$

Covariance measures how much two random variables are *linearly* related to each other (correlation). Positive covariance indicates that one variable is related to another variable linearly (e.g. when one variable is large, the other variable will also be large). The bigger the covariance is, the stronger the linear relationship becomes. Negative covariance indicates that there is an inverse linear relationship between two variables (eg. when one variable is small another variable will be big, and vice versa). When covariance is zero, X and Y do not have some linear relationship (ie. uncorrelated), but it does not necessarily imply that X and Y are independent. Again, consider continuous random variables $X \sim \text{Unif}(-1, 1)$ and $Y = X^2$. They are clearly dependent since Y depends on the value of X[5].

EXAMPLE 50.

Show that X and Y are not independent.

Independence means $P(X \le a, Y \le b) = P(X \le a)P(Y \le b)$ for any real numbers a and b. Consider $a = \frac{1}{3}$ and $b = \frac{1}{4}$.

- $X \le \frac{1}{3}$, i.e. when $-1 \le X \le \frac{1}{3}$, has probability $\frac{1 + \frac{1}{3}}{2} = \frac{2}{3}$
- $Y \le \frac{1}{4}$, i.e. when $-\frac{1}{2} \le X \le \frac{1}{2}$, has probability $\frac{1}{2}$
- $X \le \frac{1}{3}, Y \le \frac{1}{4}$ together, i.e. when $-\frac{1}{2} \le X \le \frac{1}{3}$, has probability $\frac{5}{12}$

but $\frac{2}{3} \times \frac{1}{2} \ne \frac{5}{12}$. This makes X and Y not independent.

So, we have

$$E[X] = 0$$

---

[5]Note that you can't conclude that they are independent just because $E[XY] = E[X]E[Y]$. If X and Y are independent then $E[XY] = E[X]E[Y]$ but it is not necessarily true that if $E[XY] = E[X]E[Y]$ then X and Y are independent, only imply uncorrelateness.

$$E[Y] = E[X^2] = \int_{-1}^{1} x^2 dx = \frac{2}{3}$$

$$E[XY] = E[X^3] = \int_{-1}^{1} x^3 dx = 0$$

The covariance is therefore

$$\text{Cov}(X,Y) = E[X^3] - E[X]E[X^2] = 0 - 0 \cdot \frac{2}{3} = 0$$

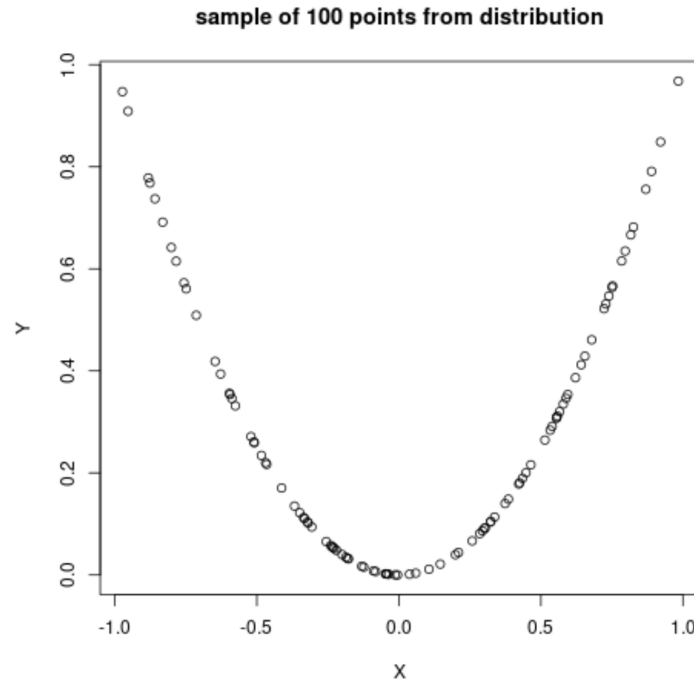This means X and Y do not depend on each other linearly.

Figure 29: Sample of 1000 points from the distribution. Points depend on each other quadratically but not linearly.

EXAMPLE 51.

Suppose 5 cards are drawn from the standard deck of 52 cards. Denote random variables X to be the number of spades in the hand and Y to be the number of black cards in the hand. Calculate $\text{Cov}(X,Y)$.

Let's try calculating $P(X = 2, Y = 4)$, the probability that a poker hand contains 4 black cards and 2 spades first. There are $\binom{52}{5}$ ways to choose 5 cards from the deck, $\binom{13}{2}$ ways to select 2 spades, $\binom{13}{2}$ ways to select the remaining 2 black cards excluding spades, and $\binom{26}{1}$ ways to choose a red card. The probability is

$$P(X = 2, Y = 4) = \frac{\binom{13}{2}\binom{13}{2}\binom{26}{1}}{\binom{52}{5}}$$

We can generalize this result to x spades and y black cards to create the joint PMF for X and Y:

$$p_{XY}(x,y) = P(X = x, Y = y) = \frac{\binom{13}{x}\binom{13}{y-x}\binom{26}{5-y}}{\binom{52}{5}}$$

This gives nonzero probability when $X \le Y$. The joint PMF table is given below:

X \ Y	0	1	2	3	4	5	
0	0.025310	0.000000	0.000000	0.000000	0.000000	0.000000	0.02531
1	0.074780	0.074780	0.000000	0.000000	0.000000	0.000000	0.14956
2	0.078031	0.169068	0.078031	0.000000	0.000000	0.000000	0.32513
3	0.035764	0.126801	0.126801	0.035764	0.000000	0.000000	0.32513
4	0.007153	0.037195	0.060864	0.037195	0.007153	0.000000	0.14956
5	0.000495	0.003576	0.008583	0.008583	0.003576	0.000495	0.02531
	0.221534	0.411420	0.274280	0.081543	0.010729	0.000495	p_Y / p_X

With expectation values

$$EX = 1.25$$

$$EY = 2.5$$

There are two ways to calculate covariance: use $E[(X - EX)(Y - EY)]$ or $E[XY] - E[X]E[Y]$. Both methods give the same answer:

$$\text{Cov}(X,Y) = \sum_{y=0}^{5}\sum_{x=0}^{5}(x - EX)(y - EY)p_{XY}(x,y) = 0.575980$$

$$\text{Cov}(X,Y) = \left(\sum_{y=0}^{5}\sum_{x=0}^{5}xyp_{XY}(x,y)\right) - (EX)(EY) = 0.575980$$

When the number of spades increases, the number of black cards also increases.

Most of the computations in this example are handled by Python[a].

---

[a]Source code

Correlation coefficient $\rho(X, Y)$ of X and Y is defined as

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

It is a unitless quantity that measures correlation between X and Y. It can be thought as the a normalized version of covariance since it takes values between -1 and 1. To prove this statement, consider the following theorem:

Theorem. *Cauchy-Schwarz Inequality:* $|E(XY)|^2 \le E(X^2)E(Y^2)$

Proof. Define a new random variable $Z = X + \lambda Y$ where λ is some real number.

The expectation for Z^2 is

$$\begin{aligned}
E(Z^2) &= E((X + \lambda Y)^2) \\
&= E(X^2 + 2\lambda XY + \lambda^2 Y^2) \\
&= E(X^2) + 2\lambda E(XY) + \lambda^2 E(Y^2)
\end{aligned}$$

Clearly, $E(Z^2)$ cannot be less than 0: $E(Z^2) \geq 0$. Let $c = E(X^2)$, $b = 2E(XY)$ and $a = E(Y^2)$,

$$a\lambda^2 + b\lambda + c \geq 0$$

Recall that a quadratic equation that gives positive values for all λ should have discriminant $b^2 - 4ac \leq 0$. This means

$$(2E(XY))^2 - 4E(X^2)E(Y^2) \leq 0$$
$$E(XY)^2 - E(X^2)E(Y^2) \leq 0$$

We finally get the equality:

$$|E(XY)|^2 \leq E(X^2)E(Y^2)$$

$\hfill\square$

We will now apply this inequality to show that $\rho(X, Y)$ is normalized.

Theorem. $\rho(X, Y)$ *only takes values between* -1 *and* 1.

Proof. We know that $\mathrm{Cov}(X, Y) = E[(X - EX)(Y - EY)]$. Applying the Cauchy-Schwarz inequality to $E[(X - EX)(Y - EX)]$ gives

$$|E[(X - EX)(Y - EY)]|^2 \leq E[(X - EX)^2]E[(Y - EY)^2]$$

By the definition of variance,

$$|\mathrm{Cov}(X, Y)|^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y)$$

$$\frac{|\mathrm{Cov}(X, Y)|^2}{\mathrm{Var}(X)\,\mathrm{Var}(Y)} \leq 1$$

Taking square root gives

$$|\rho| \leq 1$$

So $-1 \leq \rho \leq 1$. $\hfill\square$

It can be shown that

$$\rho(X, Y) = \begin{cases} 1 & \text{iff } Y = aX + b \text{ where } a, b \in \mathbb{R} \text{ and } a > 0 \\ -1 & \text{iff } Y = aX + b \text{ where } a, b \in \mathbb{R} \text{ and } a < 0 \end{cases}$$

In other words, $|\rho| = 1$ if there is a direct linear relationship between X and Y.

4.6   Variance and Independence

Theorem. *If X, Y are independent then*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Proof.

$$
\begin{aligned}
\mathrm{Var}(X + Y) &= E((X + Y)^2) - (E(X + Y))^2 \\
&= E(X^2 + 2XY + Y^2) - (EX + EY)^2 \\
&= E(X^2) + 2E(XY) + E(Y^2) - ((EX)^2 + 2(EX)(EY) + (EY)^2) \\
&= (E(X^2) - (EX)^2) + (E(Y^2) - (EY)^2) + 2(E(XY) - (EX)(EY)) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)
\end{aligned}
$$

Since X, Y are independent, $\mathrm{Cov}(X, Y) = 0$,

$$= \mathrm{Var}(X) + \mathrm{Var}(Y)$$

$\square$

This result can be generalized to multiple random variables:

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i)$$

EXAMPLE 52.

Find variance for a binomial random variable $X \sim \mathrm{Bin}(n, p)$.

We can write X in terms of n independent Bernoulli trials:

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

where X_i are Bernoulli random variables.

Since X_is are independent from each other,

$$\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i)$$

For a single trial, the variance is

$$\mathrm{Var}(X_i) = E(X_i^2) - (E(X_i))^2$$

Since $X_i^2 = X_i$ and expectation for Bernoulli random variable is p,

$$\mathrm{Var}(X_i) = E(X_i) - (E(X_i))^2 = p - p^2$$

The variance for a binomial random variable is therefore

$$\mathrm{Var}(X) = \sum_{i=1}^n p - p^2 = np(1 - p)$$

4.7   Sum of Independent Random Variables

Let X and Y be independent discrete random variables. The PMF of its sum $X + Y$ is:

$$p_{X+Y}(n) = P(X + Y = n)$$

To calculate $P(X + Y = n)$, we need to consider all possible combinations of X and Y that add up to n. This gives

$$p_{X+Y}(n) = \sum_{j=0}^{n} P(X = j, Y = n - j) = \sum_{j=0}^{n} P(X = j)P(Y = n - j)$$

where the last formula holds because X and Y are independent. The PMF $p_{X+Y}(n)$ is known as the convolution of PMFs $p_X(n)$ and $p_Y(n)$.

EXAMPLE 53.

Suppose X and Y are independent Poisson random variables with parameters λ_1 and λ_2 respectively. Find the PMF of $X + Y$.

For $n = 0, 1, 2, \ldots$, we have

$$p_{X+Y}(n) = \sum_{j=0}^{n} P(X = j)P(Y = n - j)$$

$$= \sum_{j=0}^{n} \frac{\lambda_1^j e^{-\lambda_1}}{j!} \frac{\lambda_2^{n-j} e^{-\lambda_2}}{(n-j)!}$$

$$= e^{-(\lambda_1 + \lambda_2)} \frac{1}{n!} \sum_{j=0}^{n} \binom{n}{j} \lambda_1^j \lambda_2^{n-j}$$

Applying binomial theorem $(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$,

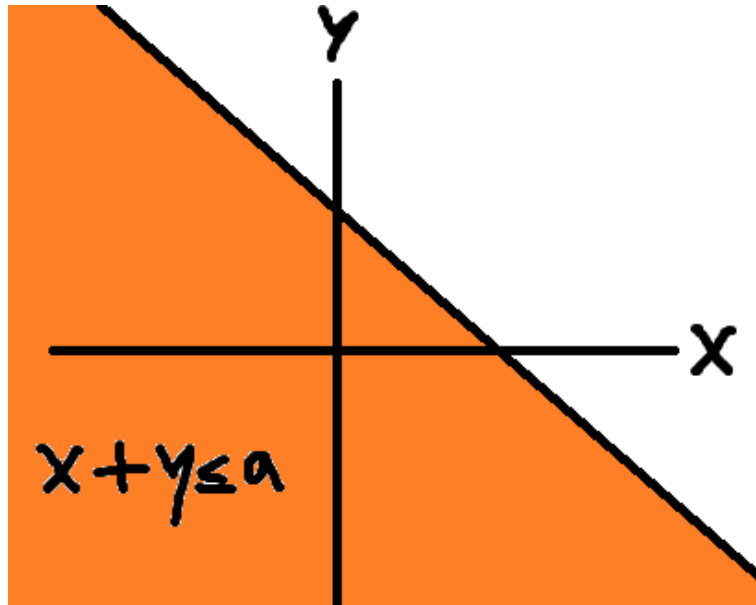$$= e^{-(\lambda_1 + \lambda_2)} \frac{(\lambda_1 + \lambda_2)^n}{n!}$$

We got $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$

The calculation is similar to binomial random variables. Suppose $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, then $X + Y \sim \text{Bin}(n + m, p)$.

Now, suppose X and Y are continuous independent random variables. The CDF for $X + Y$ is:

$$F_{X+Y}(a) = P(X + Y \leq a) = \iint_R f_X(x) f_Y(y) dx\, dy$$

where R is the region that satisfied inequality $x + y \leq a$.

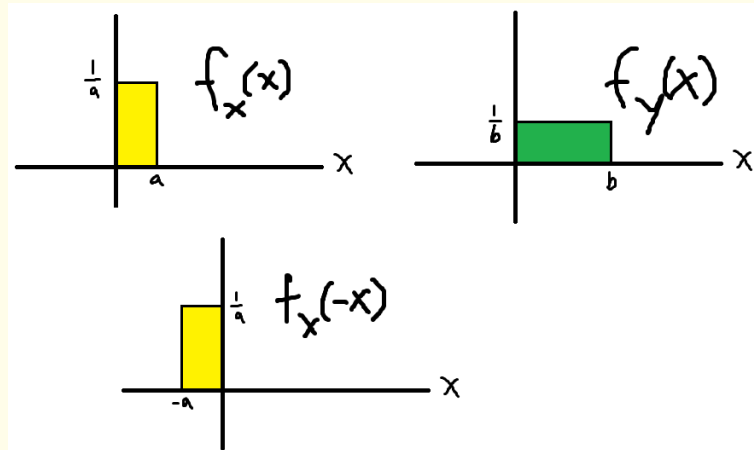Figure 30: Illustration of the Inequality $x + y \leq a$

$$F_{X+Y}(a) = \iint_R f_X(x)f_Y(y)dx\,dy$$

$$= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{a-y} f_X(x)dx \right) f_Y(y)dy$$

$$= \int_{-\infty}^{\infty} F_X(a - y)f_Y(y)dy$$

We found the convolution for $F_X(a)$ and $F_Y(a)$. To find PDF for $X + Y$, we differentiate $F_{X+Y}(a)$ which gives

$$f_{X+Y}(a) = \frac{d}{da}F_{X+Y}(a)$$

$$= \int_{-\infty}^{\infty} \left[\frac{d}{da}F_X(a - y) \right] f_Y(y)dy$$

$$= \int_{-\infty}^{\infty} f_X(a - y)f_Y(y)dy$$

EXAMPLE 54.

For $X \sim \text{Unif}(0, a)$ and $Y \sim \text{Unif}(0, b)$, calculate the probability density for $X + Y$, assuming that $a < b$.


Figure 31: Illustrations for PDFs $f_X(x)$, $f_Y(x)$ and $f_X(-x)$

The formula for finding convolution is:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(t-x)f_Y(x)dx$$

We need to integrate the product $f_X(t-x)$ (which is basically the flipped version of $f_X(x)$) and $f_Y(x)$. Unfortunately, calculating the integral is not easy. We need to consider different cases for t. The ranges of t that gives nonzero product are: $0 < t < a$, $a < t < b$, and $b < t < a + b$. Integral is different in each case.
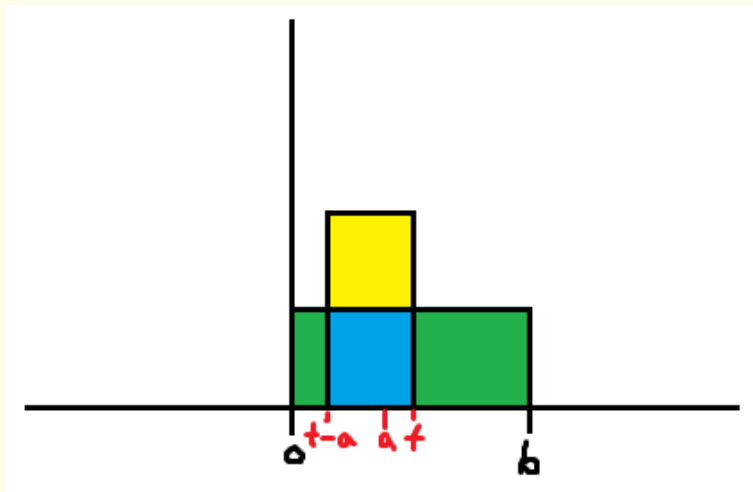
Let's start with range $0 < t < a$,
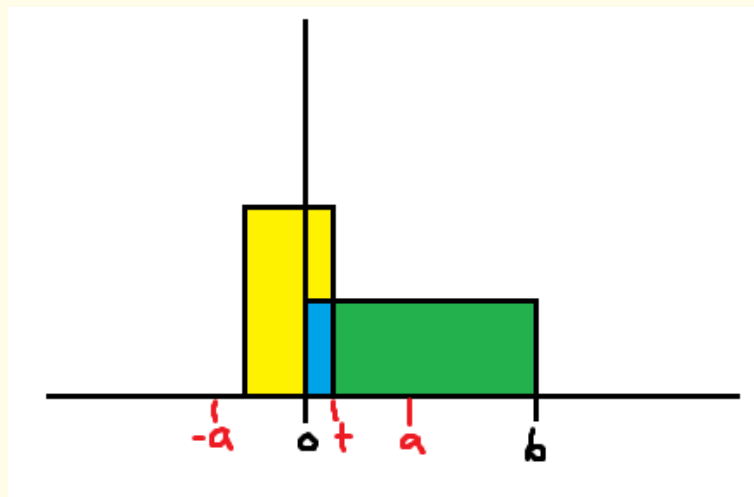
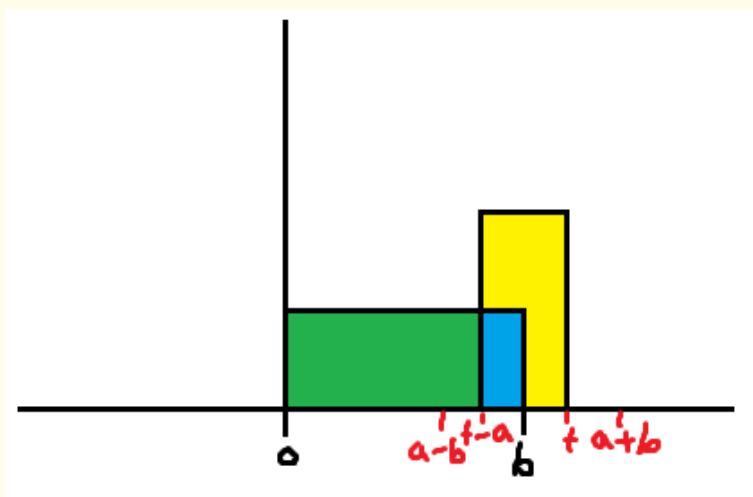The product in this range is $\frac{1}{a}\frac{1}{b} = \frac{1}{ab}$ and the integral is

$$f_{X+Y}(t) = \int_0^t \frac{1}{ab}dx = \frac{t}{ab}$$

For range $a < t < b$,

$$f_{X+Y}(t) = \int_{t-a}^{t} \frac{1}{ab} dx = \frac{1}{b}$$
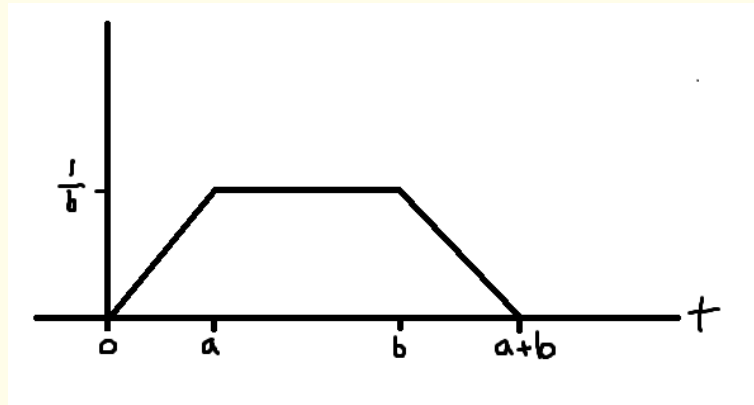
Finally for range $b < t < a + b$,

$$f_{X+Y}(t) = \int_{t-a}^{b} \frac{1}{ab} dx = \frac{a+b-t}{ab}$$

The final answer is:

$$f_{X+Y}(t) = \begin{cases} \frac{t}{ab} & 0 < t < a \\ \frac{1}{b} & a < t < b \\ \frac{a+b-t}{ab} & b < t < a + b \\ 0 & \text{otherwise} \end{cases}$$

The graph of this PDF looks like the following:

This PDF shows probable t values that can be decomposed into the sum $X + Y$. It is peaked in range $[a, b]$ because there are many ways to obtain the values between a and b than other points. There is only one way to obtain either 0 or $a + b$, so their probabilities are close to zero.

EXAMPLE 55.

Busy server queue: A web server has n pending tasks. Find a PDF for calculating total time it takes to complete all tasks.

Denote a random variable X_i to be the time taken for completing task i. They are assumed to be independent from each other and can be modelled with an exponential random variable with parameter λ.

We start by calculating PDF for $X_1 + X_2$, the time of completion of task 2. Applying the convolution formula we derived:

$$
\begin{aligned}
f_{X_1+X_2}(x) &= \int_{-\infty}^{\infty} f_{X_1}(x - y) f_{X_2}(y) dy \\
&= \int_0^x \lambda e^{-\lambda(x-y)} \lambda e^{-\lambda y} dy \\
&= \lambda^2 e^{-\lambda x} \int_0^x dy \\
&= \lambda^2 x e^{-\lambda x}
\end{aligned}
$$

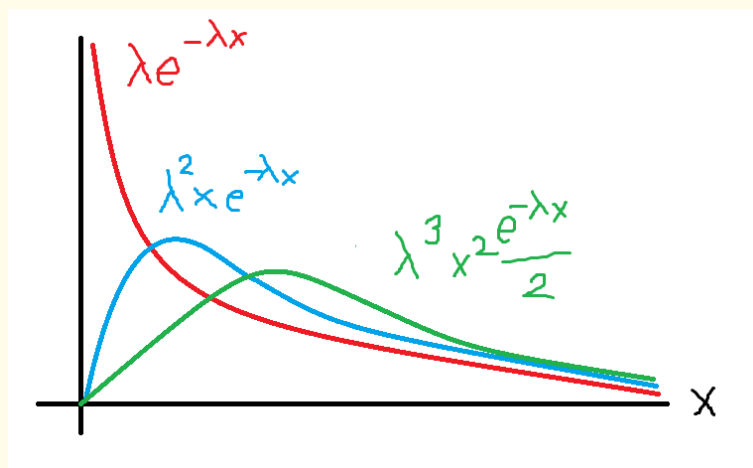The below plot compares two PDFs $f_{X_1}(x)$ and $f_{X_1+X_2}(x)$:
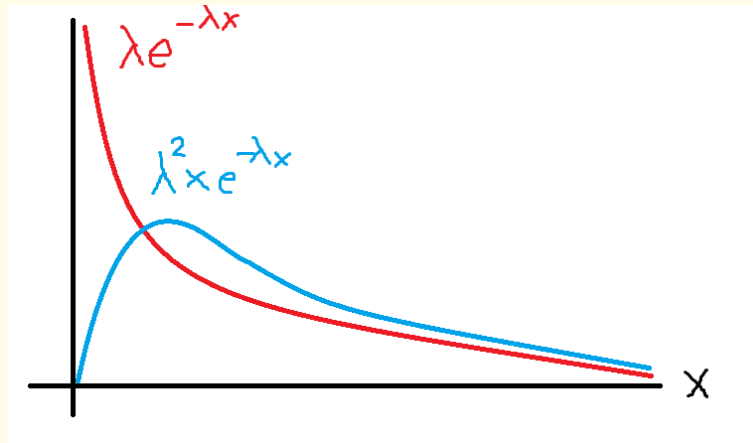
Figure 32: $f_{X_1}(x)$ and $f_{X_1+X_2}(x)$

It shows that the likely completion time for task 2 is later than the completion time for task 1.

Next, we calculate PDF for $X_1 + X_2 + X_3$ which gives:

$$f_{X_1+X_2+X_3}(x) = \int_{-\infty}^{\infty} f_{X_1+X_2}(x-y)f_{X_3}(y)dy$$
$$= \int_{0}^{x} \lambda^2(x-y)e^{-\lambda(x-y)}\lambda e^{-\lambda y}dy$$
$$= \lambda^3 x^2 \frac{e^{-\lambda x}}{2}$$

and here is the comparison again:

Figure 33: $f_{X_1}(x)$, $f_{X_1+X_2}(x)$ and $f_{X_1+X_2+X^3}(x)$

We can generalize our PDF for n tasks as follows:

$$f_{X_1+X_2+\cdots+X_n}(x) = \lambda^n x^{n-1} \frac{e^{-\lambda x}}{(n-1)!}$$

This is the distribution for Gamma random variable: $X_1+X_2+\cdots+X_n \sim \text{Gamma}(n, \lambda)$.

4.8 Teaser: Poisson Processes

The busy server queue example we did is an example of Poisson process. Task completions occur continuously and independently. Denote random variable N_t be the number of tasks completed by time t. The expectation for this random variable is:

$$E[N_t] = \lambda t$$

It makes sense because λ can be interpreted as number of tasks done per unit time (ie. rate), so multiplying it by time t will give the expected number of tasks done.

Let's find a PMF for N_t. We start by calculating $P(N_t \geq n)$ which is equivalent to $P(X_1 + X_2 + \cdots + X_n \leq t)$, the probability that n tasks completed by time t.

$$P(N_t \geq n) = P(X_1 + X_2 + \cdots + X_n \leq t)$$
$$= \int_0^t \frac{\lambda^n x^{n-1}}{(n-1)!} e^{-\lambda x} dx$$

Change of variables using $u = \lambda x$ gives,

$$= \int_0^{\lambda t} \frac{u^{n-1}}{(n-1)!} e^{-u} du$$

Integration by parts gives,

$$= -\frac{u^{n-1}}{(n-1)!} e^{-u} \Big|_0^{\lambda t} + \int_0^{\lambda t} \frac{u^{n-2}}{(n-2)!} e^{-u} du$$
$$= -\frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t} + P(N_t \geq n-1)$$

Finally,

$$P(N_t = n-1) = P(N_t \geq n-1) - P(N_t \geq n) = \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t}$$

We conclude that N_t is a Poisson random variable with parameter λt: $N_t \sim \text{Pois}(\lambda t)$.

5   Generating Functions and Convergence

5.1   Moment Generating Functions

The moment generating function (MGF) of a random variable X is defined as:

$$M_X(t) = E[e^{tX}] = \begin{cases} \displaystyle\int_{-\infty}^{\infty} e^{tx} f_X(x)dx & X \text{ continuous} \\[2em] \displaystyle\sum_i e^{tx} p_X(x_i) & X \text{ discrete} \end{cases}$$

for all t values. As its name suggests, this function can be used to calculate moments of arbitrary order: EX, $E[X^2]$, $E[X^3]$, etc.. To see why this is true, we apply Taylor expansion on e^{tX}:

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

So to get nth moment of X, we simply need to differentiate its MGF n times then set $t = 0$:

$$E[X^n] = \frac{d^n}{dt^n}M_X(t)\bigg|_{t=0}$$

Here is a nice property for MGFs:

Theorem. *If X and Y are independent random variables,*

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

Proof. Recall that for independent X, Y, $E[XY] = E[X]E[Y]$.

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$$

$\square$

Another important property of MGFs is that it can uniquely determine the distribution of X.

Lastly, one thing to keep in mind is that MGFs does not always exist. For example, consider a continuous random variable $X \sim \text{Exp}(\lambda)$. Its MGF is:

$$M_X(t) = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x}dx = \int_0^{\infty} \lambda e^{(t-\lambda)x}dx = \begin{cases} \infty & t \geq \lambda \\ \frac{\lambda}{\lambda-t} & t < \lambda \end{cases}$$

EXAMPLE 56.

For $X \sim \text{Pois}(\lambda)$, find its MGF and use it to find $E[X]$.

$$M_X(t) = E[e^{tX}]$$

$$= \sum_{n=0}^{\infty} e^{tn} e^{-\lambda} \frac{\lambda^n}{n!}$$

$$= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!}$$

$$= e^{-\lambda} e^{\lambda e^t}$$

$$= \exp(\lambda(e^t - 1))$$

To get the first moment, we differentiate its MGF once and setting $t = 0$:

$$E[X] = \frac{d}{dt} \left[\exp(\lambda(e^t - 1)) \right]\Big|_{t=0} = \lambda e^t \exp(\lambda(e^t - 1))\big|_{t=0} = \lambda$$

5.2 Characteristic Functions

The characteristic function of a random variable X is defined as:

$$\phi_X(t) = M_X(it) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} e^{itx} f_X(x) dx & X \text{ continuous} \\[2em] \displaystyle\sum_i e^{itx} p_X(x_i) & X \text{ discrete} \end{cases}$$

They are essentially Fourier transforms of PDF/PMF with frequency denoted by t.

Its properties are:

i $\phi(t)$ always exists and it obeys $|\phi(t)| \leq 1$[6].

ii $\phi(0) = E[e^{i0X}] = 1$

iii If random variables X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

iv Uniformly continuous everywhere.

Characteristic functions encode every piece of information about a random variable, including its moments and distribution.

To calculate nth moment of a random variable X using its characteristic function $\phi_X(t)$, observe that

$$\frac{d^n}{dt^n} \phi_X(t)\Big|_{t=0} = E\left[\frac{d^n}{dt^n} e^{itX}\Big|_{t=0} \right] = E[(iX)^n]$$

---

[6]To see it more explicitly, $\left| \int e^{itx} f(x) dx \right| \leq \int \underbrace{|e^{itx}|}_{=1} f(x) dx = 1$. Recall that $|e^{i\theta}| = \cos^2\theta + \sin^2\theta = 1$ and

$\int f(x) dx = 1$.

So,

$$E[X^n] = \frac{1}{i^n} \frac{d^n}{dt^n} \phi_X(t) \Big|_{t=0}$$

From $\phi_X(t)$, we can recover the distribution of a random variable X.

Theorem. *If X is continuous with PDF $f_X(x)$, then $f_X(x)$ can be recovered by performing inverse Fourier transform on its characteristic function $\phi_X(t)$:*

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t) dt$$

for every X where $f_X(x)$ exists.

This theorem can be proved with Fourier inversion theorem which is beyond the scope of this course.

EXAMPLE 57.

Find the characteristic function for $X \sim \text{Exp}(\lambda)$ and use it to find $f_X(0)$.

$$\phi_X(t) = \int_0^{\infty} e^{itx} \lambda e^{-\lambda x} dx$$
$$= \lambda \int_0^{\infty} e^{-(\lambda - it)x} dx$$
$$= \frac{\lambda}{\lambda - it}$$

We apply inverse Fourier transform to find $f(0)$,

$$f(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\lambda}{\lambda - it} dt$$
$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{1 - \frac{it}{\lambda}} dt$$

Change of variables using $s = \frac{t}{\lambda}$ gives:

$$= \frac{\lambda}{2\pi} \int_{-\infty}^{\infty} \frac{1}{1 - is} ds$$

By some black integration magic...

$$= \frac{\lambda}{2}$$

Recall that for an exponential random variable, PDF $f(x)$ at $x = 0$ either takes 0 or λ. This result is the average between two values:

$$f(0) = \frac{f(0^+) + f(0^-)}{2} = \frac{\lambda + 0}{2} = \frac{\lambda}{2}$$

Theorem. *For a random variable X and real numbers a and b,*

$$\phi_{aX+b}(t) = e^{itb}\phi_X(at)$$

Proof.

$$\phi_{aX+b}(t) = E[e^{it(aX+b)}] = e^{itb}E[e^{i(ta)X}] = e^{itb}\phi_X(at)$$

$\square$

In particular, if $X = $ c where c is some constant then $\phi(t) = e^{itc}$.

EXAMPLE 58.

Find characteristic functions for $X \sim \text{Norm}(0,1)$ and $Y \sim \text{Norm}(\mu, \sigma^2)$.

Let's find the characteristic function for standard normal:

$$\phi_X(t) = E[e^{itX}]$$
$$= \int_{-\infty}^{\infty} e^{itx}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx$$

Recall $e^{itx} = \cos(tx) + i\sin(tx)$. sin integrates to 0 since it is an odd function, so

$$= \int_{-\infty}^{\infty} \cos(tx)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx$$

This is a Gaussian integral. By looking up the table of integrals,

$$= \exp\left(-\frac{t^2}{2}\right)$$

For Y, we can rescale it to standard normal by using this substitution $Z = \frac{Y-\mu}{\sigma}$. Invoking the theorem we introduced earlier:

$$\phi_Y(t) = \phi_{\sigma Z+\mu}(t) = e^{it\mu}\phi_Z(\sigma t) = e^{it\mu}\exp\left(-\frac{\sigma^2 t^2}{2}\right)$$

EXAMPLE 59.

For n independent and identitcally distributed (iid.) exponential random variables: $X_1, X_2, \ldots, X_n \sim \text{Exp}(\lambda)$, denote the its sum $S_n = X_1 + X_2 + \cdots + X_n$ and its average $\frac{S_n}{n}$. Find the characteristic function for the average of n exponential random variables. What happens when $n \to \infty$?

We know that

$$\phi_{X_j}(t) = \frac{\lambda}{\lambda - it}$$

for all j from 1 to n. The characteristic function for S_n is:

$$\phi_{S_n}(t) = \phi_{X_1+X_2+\cdots+X_n}(t) = \left(\frac{\lambda}{\lambda - it}\right)^n$$

since X_j are independent from each other. Then,

$$\phi_{\frac{S_n}{n}}(t) = \phi_{S_n}\left(\frac{t}{n}\right) = \left(\frac{\lambda}{\lambda - \frac{it}{n}}\right)^n$$

Now as $n \to \infty$,

$$\phi_{\frac{S_n}{n}}(t) = \lim_{n \to \infty}\left(\frac{1}{1 - \frac{it}{\lambda n}}\right)^n = \frac{1}{e^{-\frac{it}{\lambda}}} = e^{\frac{it}{\lambda}}$$

We see that $e^{\frac{it}{\lambda}}$ is the characteristic function of constant $\frac{1}{\lambda}$.

EXAMPLE 60.

Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$ be independent binomial random variables. Show that $X + Y \sim \text{Bin}(n + m, p)$.

The characteristic function for X is

$$\phi_X(t) = (1 - p + e^{it})^n$$

and for Y

$$\phi_X(t) = (1 - p + e^{it})^m$$

Since X and Y are independent, we have

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) = (1 - p + e^{it})^n(1 - p + e^{it})^m = (1 - p + e^{it})^{n+m}$$

This gives the characteristic function for $\text{Bin}(n+m, p)$, therefore $X + Y \sim \text{Bin}(n+m, p)$.

Remark: One can also use MGFs to prove this.

5.3 Convergence of Distributions

Let $F_1, F_2, \ldots, F_n$ denote CDF of some random variables $X_1, X_2, \ldots, X_n$. Also, $F_n \Rightarrow F$ is defined as

$$\lim_{n \to \infty} F_n(x) = F$$

whenever F is continuous at x.

EXAMPLE 61.

Continuing from the previous example, The PDF and CDF for $\frac{S_n}{n}$ looks like the following:
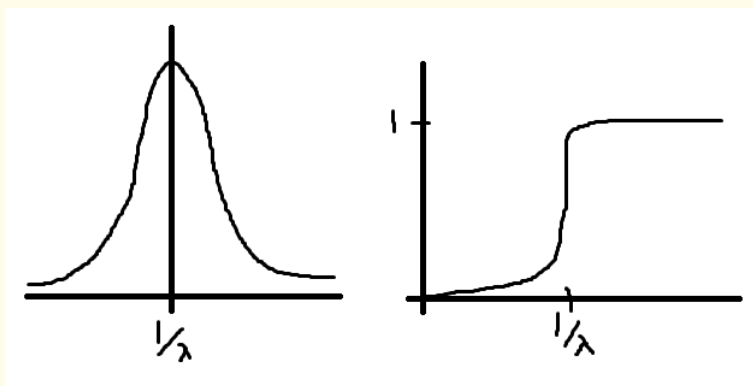

Figure 34: PDF and CDF for $\frac{S_n}{n}$

As $n \to \infty$, we see that

$$
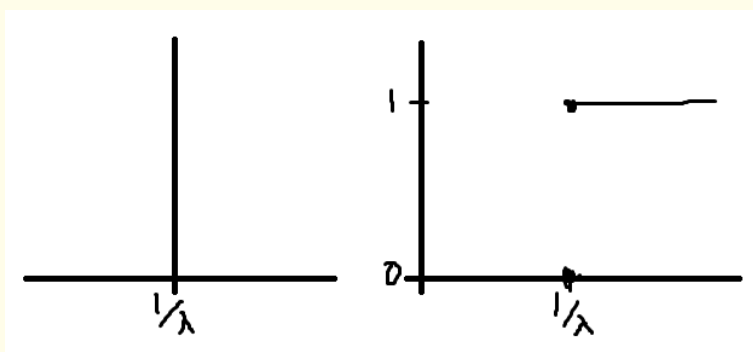\lim_{n \to \infty} F_n(x) = \begin{cases} 0 & x < 1/\lambda \\ 1 & x > 1/\lambda \end{cases}
$$

Figure 35: PDF and CDF for $\frac{S_n}{n}$ as $n \to \infty$

EXAMPLE 62.

Suppose a random variable X_n is uniform on $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$. As $n \to \infty$, its CDF will look like the CDF for $\text{Unif}(0,1)$.
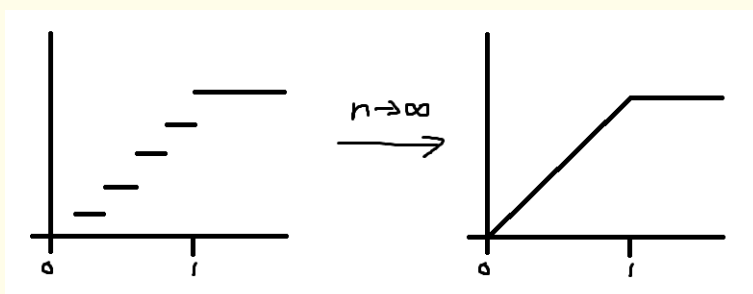

Figure 36: CDFs for X_n

If a random variables X_n with CDF F_n have $F_n \Rightarrow F$, we say that X_n *converge* in distribution:

$$
X_n \xrightarrow{D} X
$$

if X has CDF F. In other words, $F_{X_n}(x) \Rightarrow F_X(x)$ whenever F_X continuous.

Theorem. *Continuity Theorem: If random variables X_n with CDF F_n and characteristic function $\phi_n(t)$, then $\phi_n(t) \to \phi(t)$. If $\phi_n(t) \to \phi(t)$ for all t and $\phi(t)$ is continuous at 0, then $\phi(t)$ is the characteristic function of X and $X_n \xrightarrow{D} X$.*

EXAMPLE 63.

Continuing from the $\frac{S_n}{n}$ example, we saw that

$$\lim_{n \to \infty} \phi_{\frac{S_n}{n}}(t) = e^{\frac{it}{\lambda}}$$

By continuity theorem,

$$\frac{S_n}{n} \xrightarrow{D} \frac{1}{\lambda}$$

That is, the CDF of $\frac{S_n}{n}$ approximates to CDF of constant $\frac{1}{\lambda}$ (see figure 34).

So, $P(a < \frac{S_n}{n} < b) = P(a < \frac{1}{\lambda} < b)$. In other words, the average of $X_1, X_2, \ldots, X_n$ is approximately $\frac{1}{\lambda}$ with high probability.

Theorem. *Weak Law of Large Numbers: Let $X_1, X_2, \ldots, X_n$ be iid. random variables with the same distribution. Let the expectation for each random variable X_i, $\mu = EX_i$ is finite. Let $S_n = \sum_{i=1}^{n} X_i$, then*

$$\frac{S_n}{n} \xrightarrow{D} \mu$$

Proof. By continuity theorem, it suffices to show that

$$\lim_{n \to \infty} \phi_{\frac{S_n}{n}}(t) = e^{it\mu}$$

for all t. We know that

$$\phi_{\frac{S_n}{n}}(t) = \left(\phi_X \left(\frac{t}{n} \right) \right)^n$$

Applying Taylor's expansion on ϕ_X about $t = 0$ gives:

$$\phi_X(s) = 1 + s\phi_X'(0) + O(s^2) = 1 + i(EX)s + O(s^2)$$

Substituting this result into $\phi_X \left(\frac{t}{n} \right)$ gives

$$\phi_X \left(\frac{t}{n} \right) = 1 + \frac{it\mu}{n} + O \left(\frac{t^2}{n^2} \right)$$

Finally,

$$\lim_{n \to \infty} \phi_{\frac{S_n}{n}}(t) = \lim_{n \to \infty} \left(1 + \frac{it\mu}{n} + O \left(\frac{t^2}{n^2} \right) \right)^n = e^{it\mu}$$

$\square$

Note that the continuity theorem needs the assumption that ϕ is continuous at 0 otherwise ϕ will not be characteristic function of any random variable. As an example, consider the following characteristic function for $\text{Unif}(-n, n)$:

$$\phi_n(t) = \begin{cases} \frac{\sin(nt)}{nt} & t \neq 0 \\ 1 & t = 0 \end{cases}$$

As $n \to \infty$,

$$\phi_n(t) = \begin{cases} 0 & t \neq 0 \\ 1 & t = 0 \end{cases}$$

This is not the characteristic function since it is discontinuous.

5.4 Central Limit Theorem

Consider an experiment where a fair coin is tossed n times. The number of heads can be modelled with a binomial random variable $X \sim \text{Bin}(n, 1/2)$. Its PMF is

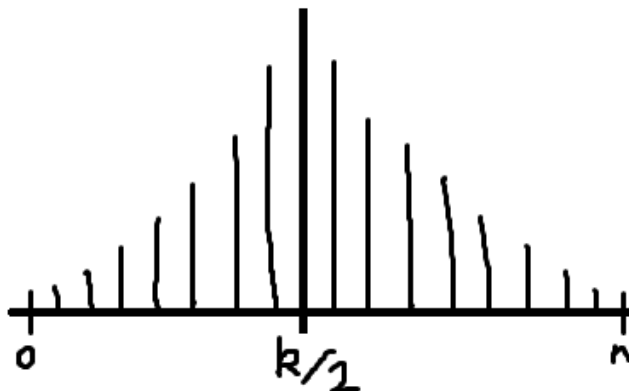$$p(k) = \binom{n}{k} \left(\frac{1}{2}\right)^n$$

Figure 37: Distribution for Coin Tosses

The distribution looks almost like a bell curve. Let's see if we can transform the PMF into normal PDF. We start by applying Stirling's approximation $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$.

$$p(k) = \frac{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n}{\sqrt{2\pi k} \left(\frac{k}{e}\right)^k \sqrt{2\pi(n-k)} \left(\frac{n-k}{e}\right)^{n-k} 2^n}$$

When $k = \frac{n}{2}$,

$$p\left(\frac{n}{2}\right) \approx \frac{1}{\sqrt{\pi n}}$$

79

We introduce a parameter t which denotes variation from $\frac{n}{2}$. So when $k = \frac{n}{2} + t$,

$$p\left(\frac{n}{2} + t\right) \approx \frac{1}{\sqrt{\pi n}} \exp\left(-\frac{t^2}{2n}\right)$$

This is a normal approximation which is valid when $n \gg 1$ and $t \ll n$. Poisson random variable also has the similar approximation:

$$p(\lambda + t) = \frac{1}{\sqrt{2\pi\lambda}} \exp\left(-\frac{t^2}{2\lambda}\right)$$

Theorem. *Central Limit Theorem (CLT): Let $X_1, X_2, \ldots, X_n$ be iid. random variables, each with finite expectation $\mu = E[X_i]$ and finite variance $\sigma^2 = \text{Var}(X_i)$. Let $\bar{X}_n = \frac{1}{n}\sum_i X_i$ be the sample mean and $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ or $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ where $S_n = X_1 + X_2 + \cdots + X_n$. Then we have*

$$Z_n \xrightarrow{D} \text{Norm}(0, 1)$$

That is, as $n \to \infty$,

$$P\left(a \le Z_n \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(-\frac{x^2}{2}\right) dx$$

for all a and b.

Proof. We start by defining a new random variable $Y_i = \frac{X_i - \mu}{\sigma}$ such that

$$\sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}} = Z_n$$

Y_i has zero expectation $E[Y_i] = 0$ and unit variance $\text{Var}(Y_i) = 1$. By continuity theorem, it suffices to show that

$$\phi_{Z_n}(t) \to e^{-\frac{t^2}{2}}$$

as $n \to \infty$.

$$\phi_{Z_n}(t) = \phi_{Y_1}\left(\frac{t}{\sqrt{n}}\right) \phi_{Y_2}\left(\frac{t}{\sqrt{n}}\right) \cdots \phi_{Y_n}\left(\frac{t}{\sqrt{n}}\right) = \left[\phi_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

Applying Taylor's expansion on $\phi_{Y_i}\left(\frac{t}{\sqrt{n}}\right)$ gives:

$$\phi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{it}{\sqrt{n}}E[Y_i] + \frac{i^2 t^2}{2n}E[Y_i^2] + O(t^3) = 1 - \frac{t^2}{2n} + O(t^3)$$

Finally,

$$\lim_{n\to\infty} \phi_{Z_n}(t) = \lim_{n\to\infty} \left[1 - \frac{t^2}{2n} + O(t^3)\right]^n = e^{-\frac{t^2}{2}}$$

$\square$

Remarks: If $EX_i = \infty$ or not defined then $\bar{X}_n$ has no limit. If EX_i is defined but $\text{Var}(X_i) = \infty$ then $\bar{X}_n \to \mu$ but $|\bar{X}_n - \mu| \gg 1/\sqrt{n}$.

EXAMPLE 64.

UBC wants to have 50 new students in a certain program this year. Historically, only 40% of admitted students accept the offer. If 100 students admitted, what is the

probability that more than 50 people accept the offer?

We can model the number of students who accepted the offer using a binomial random variable $X \sim \text{Bin}(100, 0.4)$. We need to find $P(X > 50)$.

We observe that X is the sum of 100 Bernoulli random variables with paramater $p = 0.4$. The mean and variance for each Bernoulli random variable are $\mu = p = 0.4$ and $\sigma^2 = p(1 - p) = 0.24$.

We define a new random variable Z:

$$Z = \frac{X - np}{\sqrt{np(1 - p)}} = \frac{X - 40}{\sqrt{24}}$$

By CLT, Z can be approximated using standard normal $\text{Norm}(0, 1)$. So,

$$P(X > 50) = P\left(Z > \frac{50 - 40}{\sqrt{24}}\right) = P(Z > 2.04) = 1 - \phi(2.04)$$

Appendix A   Table of Distributions

RV	PMF/PDF	Mean	Variance	Char. Func.
$\text{Bern}(p)$	$p(n) = p^n(1-p)^{1-n}$	p	$p(1-p)$	$pe^{it} + (1-p)$
$\text{Bin}(n,p)$	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$	$(1 - p + e^{it})^n$
$\text{Geom}(q)$	$p(k) = (1-q)^{k-1}q$	$1/q$	$\dfrac{1-q}{q^2}$	$\dfrac{qe^{it}}{1 - (1-q)e^{it}}$
$\text{Pois}(\lambda)$	$p(n) = e^{-\lambda}\dfrac{\lambda^n}{n!}$	λ	λ	$\exp\left(\lambda(e^{it}-1)\right)$
$\text{Unif}(a,b)$	$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$	$\dfrac{e^{ita} - e^{itb}}{it(b-a)}$
$\text{Exp}(\lambda)$	$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$	$1/\lambda$	$1/\lambda^2$	$\dfrac{\lambda}{\lambda - it}$
$\text{Norm}(\mu, \sigma^2)$	$f(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2	$\exp\left(it\mu - \dfrac{\sigma^2 t^2}{2}\right)$