May 3, 2023

Probability course notes

by

Fred Torcaso

CONTENTS.

* can be skipped on first reading.

These notes were developed over five semesters (Spring 2021 through Spring 2023) for the course *Introduction to Probability* at the Johns Hopkins University. This course is a comprehensive treatment of applied probability taught at the senior level.

I. Probability at the experiment level and Combinatorics.

The point of this section is to carefully and rigorously explain the mathematics of counting in a systematic manner. Equal emphasis will be on modeling, counting and computing probability in the case of experiments leading to a finite number of equally likely outcomes. Many examples are provided.

In future sections we will need to come back to these roots. We will also later develop tools that, when combined with the combinatorics we learn here, can greatly increase our abilities to compute probability at the experiment level.

Experiment level probability.

An ***experiment*** is a repeatable process of observation where every possible outcome is known in advance.

The individual outcomes that can happen are called ***sample points***. We denote a generic sample point by the lower-case Greek letter omega ω.

The collection of all sample points is called the ***sample space*** denoted by Ω. Thus, $\omega \in \Omega$ means ω is one of the possible outcomes (sample points) of the experiment.

Basic experiments:
Please understand that in the examples below specific choices were made to *model* the sample spaces, i.e., as a way to ***specify*** what the experimenter observes as the outcome.

1.1. Flip a coin once: $\Omega = \{h, t\}$.
1.2. Flip a coin twice: $\Omega = \{hh, ht, th, tt\}$.
1.3. Flip a coin thrice: $\Omega = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$, etc.
1.*n*. Flip a coin $n > 1$ times: $\Omega = \{x_1 x_2 \cdots x_n : x_i \in \{h, t\} \text{ for all } i\}$.
We may think of these experiments as repeated trials of flipping a coin once.

2.1. Roll a single 6-sided die: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
2.2. Roll a 6-sided die twice (xy means x on 1st roll, y on 2nd roll):

$$\begin{aligned}\Omega = \{&11, \ 12, \ 13, \ 14, \ 15, \ 16,\\ &21, \ 22, \ 23, \ 24, \ 25, \ 26,\\ &31, \ 32, \ 33, \ 34, \ 35, \ 36,\\ &41, \ 42, \ 43, \ 44, \ 45, \ 46,\\ &51, \ 52, \ 53, \ 54, \ 55, \ 56,\\ &61, \ 62, \ 63, \ 64, \ 65, \ 66\}\end{aligned}$$

Again, we may wish to think of this experiment as two trials of rolling the die once.

3.1. There are 4 slips of paper (numbered 1,2,3,4) in a hat. The experiment is to draw all slips of paper one at a time noting the number on each.

$$\begin{aligned}\Omega = \{&1234, \ 1243, \ 1324, \ 1342, \ 1423, \ 1432,\\ &2134, \ 2143, \ 2314, \ 2341, \ 2413, \ 2431,\\ &3124, \ 3142, \ 3214, \ 3241, \ 3412, \ 3421,\\ &4123, \ 4132, \ 4213, \ 4231, \ 4312, \ 4321\}.\end{aligned}$$

3.2. There are 4 slips of paper (numbered 1,2,3, and 4) in a hat. The experiment is to draw *two* slips of paper one at a time noting the number on each.

$$\begin{aligned}\Omega = \{&12, \ 13, \ 14,\\ &21, \ 23, \ 24,\\ &31, \ 32, \ 34,\\ &41, \ 42, \ 43\}.\end{aligned}$$

4. There are 4 slips of paper (numbered 1,2,3, and 4) in a hat. The experiment is to draw *two* slips of paper at once noting which two you selected. We neglect order.

$$\Omega = \Big\{\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}\Big\}.$$

An ***event*** is a subset of Ω. We usually think of an event as a particular subcollection of Ω containing sample points that are all of interest to the experimenter.

When an experiment is performed, an $\omega \in \Omega$ is produced. If A is an event, we say A ***occurs*** to mean $\omega \in A$, i.e., the ω that chance produced is a member of A.

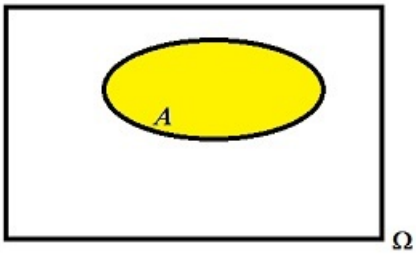


Figure. A Venn diagram visualization of an event A in Ω.

We imagine the box (Ω) filled with all its sample points. The event in yellow (A) corrals a certain group of these sample points.

Some interesting events:

In experiment 1.3, the event H_i the ith toss is a head:
$H_1 = \{hhh, hht, hth, htt\}$ $H_2 = \{hhh, hht, thh, tht\}$ $H_3 = \{hhh, hth, thh, tth\}$;
The event D the first and last toss differ in parity: $D = \{hht, htt, thh, tth\}$.

In experiment 2.2, the event S_7 the sum total is 7: $S_7 = \{16, 25, 34, 43, 52, 61\}$;
The event A a six on first roll: $A = \{61, 62, 63, 64, 65, 66\}$;
The event* B a six is rolled: $B = \{16, 26, 36, 46, 56, 66, 61, 62, 63, 64, 65\}$;
*Notice in this example that a six is rolled means *at least* one six is rolled.
The event C_1 the red die (or the first die) shows a six: $C_1 = \{61, 62, 63, 64, 65, 66\}$.
The event C_2 the green die (or the second die) shows a six: $C_1 = \{16, 26, 36, 46, 56, 66\}$.

In experiment 3.1, the event I that even numbers are in increasing order and odd numbers are in increasing order :
$I = \{1234, 1243, 1324, 2134, 2143, 2413\}$.

In experiment 3.2, the event T the sum of the numbers drawn is divisible by 3:
$B = \{12, 21, 24, 42\}$.

In experiment 4, the event F the number 4 is selected: $T = \Big\{\{1, 4\}, \{2, 4\}, \{3, 4\}\Big\}$.

Computing probabilities.

When given an experiment, we may want to understand how likely a particular event is to occur. That is, if A is an event, then we wish to assign a number $P(A)$ that tells us how likely the event is. We will learn that the computation of $P(A)$ will depend on the experiment and the probability model chosen for it. More on this later. But, for now, we introduce a historically significant way to compute $P(A)$ under the assumptions that the sample space is comprised of ***finite and equally-likely outcomes***.

The classical probability measure.
Under the assumption that $|\Omega|$ is finite and all $\omega \in \Omega$ have the same chance of occurring, then for any event $A \subseteq \Omega$,

$$P(A) = \frac{|A|}{|\Omega|}.$$

That is, $P(A)$ is just the proportion of possible outcomes that belong to A. Therefore, in the case of finite and equally-likely outcomes, computing $P(A)$ boils down to just two counting problems: count the number of things in Ω, count the number of things in A, and take the ratio.

In most problems that we will encounter we will want to count things by *avoiding* an actual count. The secret to counting effectively will be in how we *model* our sample points and how we *think* about them. The mathematical discipline of counting without actually doing a count is called ***combinatorics***.

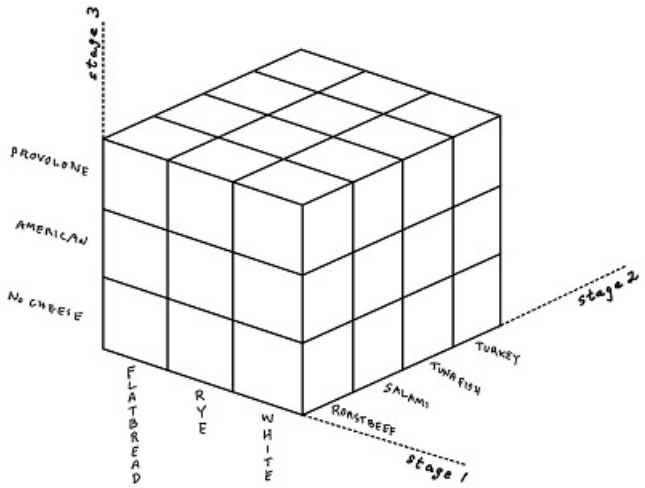
An important step to counting the number of sample points is to mathematically model your experiment that makes counting sample points *easier*; for instance, in situations when we can model our experiment as a (finite) sequence of two or more stages, where on each stage we know the number of ways that stage can be completed.

The basic counting rule.
Suppose we have a procedure consisting of r stages numbered 1 thru r. Stage 1 can be performed in n_1 ways, and, for each $k = 2, \ldots, r$, stage k can be completed in n_k ways regardless of the way stage j was completed for all $j < k$. Then the procedure can be completed in $n_1 \times n_2 \times \cdots \times n_r$ ways.

Basic example 1.
Suppose at a certain deli a *sandwich* is

(1) a choice of bread (one of Flatbread, Rye, or White),
(2) a choice of protein (one of Roast beef, Salami, Tuna fish, or Turkey), and
(3) a choice of cheese (one of no cheese, American, or Provolone)

There are 3 stages to make a sandwich. There are $n_1 = 3$ ways we can choose the bread. There are $n_2 = 4$ ways to choose the protein. There are $n_3 = 3$ ways to choose the cheese. The number of choices at each stage does not depend on *which* choice was made in previous stages. Then, there are $3 \times 4 \times 3 = 36$ possible sandwiches.



Pictorially, we can think of each box as an ordered 3-tuple (x, y, z), where $x \in \{\text{Flatbread, Rye, White}\}$, $y \in \{\text{Roast beef, Salami, Tuna Fish, Turkey}\}$, and $z \in \{\text{No cheese, American, Provolone}\}$.

We can also visualize all sandwiches through a ***tree diagram*** (see next page).

The tree below has the property that all nodes within a fixed stage have the same number of arcs leaving it. This is the paradigm of the basic counting rule.

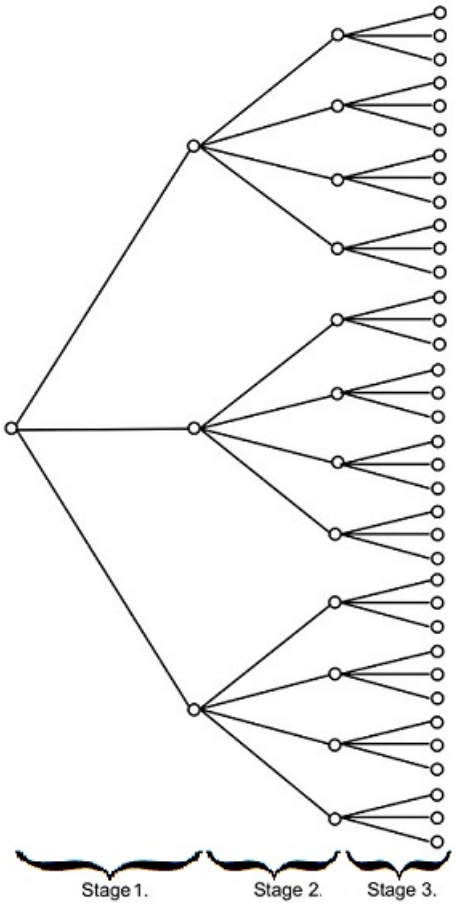


Figure. A tree diagram of all possible sandwiches.

Each path from the root node to a leaf represents one sandwich. Therefore, the number of leaves *is* the number of sandwiches.

(Stage 1) The root has 3 children.
(Stage 2) Each of these children has 4 children.
(Stage 3) Each of these children have 3 children (the leaves).
So, there are $3 \times 4 \times 3$ leaves (sandwiches).

The next example sets up a situation where the basic counting rule cannot be applied.

Basic example 2.
Suppose 5 people Amir, Betty, Cristoff, Dahlia, and Ernesto are senior members of a club from which we need to elect a (1) president, (2) a vice-president, and (3) a treasurer. No one can hold more than one office. How many elections are possible?

An election result is a choice of president, vice-president, and treasurer. Therefore, we think of the procedure as 3 stages: stage 1 (say, choose a president) can be completed in 5 ways. Stage 2 (say, choose a vice-president) can be completed in 4 ways - note that this stage can always be completed in 4 ways *regardless* of which person was chosen in the first stage; finally, stage 3 (choose a treasurer) can always be completed in 3 ways once we've used up the two people for president and vice-president. Therefore, there are $5 \times 4 \times 3 = 60$ possible election results. See figure below.

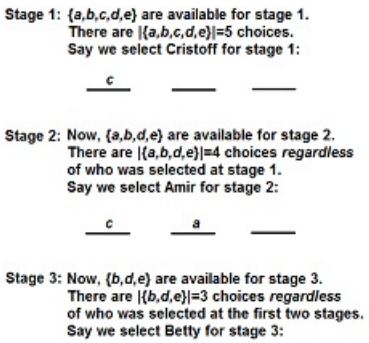


Figure. An illustration of the basic counting rule generating a sample point.

Remark. (A warning about blindly using the basic counting rule.)
Suppose that Amir refuses to hold office if Betty is elected president. How many election results are possible now? Although there may be 5 choices for president, the number of choices for vice-president (and treasurer) depends on whether or not Betty was elected president (if Betty is elected president, then there are only 3 choices for vice-president: *we must remove* Amir; if Betty is not elected president, there are 4 choices for vice-president now). So, strictly speaking, this question does not fit the paradigm of the basic counting rule: the number of choices in a stage *depends* on how a previous stage was completed.

Nevertheless, if Betty is elected president (one way to complete this), then there are 3 choices for vice-president and then 2 choices for treasurer as we must remove Amir from the pool. If Betty is not elected president (4 ways to complete this), then there are any of 4 people to be vice-president (Betty doesn't mind holding office with Amir), and then 3 ways to elect a treasurer. Thus, $1 \times 3 \times 2 + 4 \times 4 \times 3 = 54$ election results.

Remark.
Had we modeled the procedure as choosing a treasurer in stage 1, a vice-president in stage 2, and a president in stage 3, then, of course, there are still $5 \times 4 \times 3 = 60$ unrestricted election results; the only difference is the ordered 3-tuple (x, y, z) has x as *treasurer* instead of president like before, y as vice-president, and z now as the president. Having chosen to model the procedure in this fashion might have made answering the question in the previous remark more difficult; however, it can make answering the following question easier:

> Suppose that Dahlia refuses to be vice-president if Ernesto is treasurer (but all other possibilities are fine). How many election results?

With this alternate model for the procedure, when Ernesto is chosen as treasurer, there are only 3 choices of vice-president (Dahlia must be removed from the pool) and there are 3 choices for president (as Dahlia is okay with the presidency when Ernesto is the treasurer); on the other hand, if Ernesto is not treasurer (4 ways to complete this), then there are 4 ways to choose the vice-president and 3 ways to choose the president. Therefore, there are $1 \times 3 \times 3 + 4 \times 4 \times 3 = 57$ possible election results.

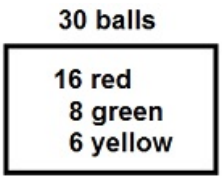
Perhaps an easier way to have solved the last problem was to recognize that out of the total $5 \times 4 \times 3 = 60$ unrestricted election results we just need to remove all those ordered 3-tuples of the form (e, d, z), where e is Ernesto, d is Dahlia, and z can be any of the 3 remaining people as president, and clearly there are 3 such "bad" election results. So, we remove these 3 bad ones from the 60: $60 - 3 = 57$.

Advice:
The choice of model can help simplify an approach to a problem. It is always beneficial to visualize yourself actually performing the experiment! The specific questions can help us choose the model we want to use to answer these questions.

Example and discussion.
We have an urn with 30 balls: 16 are red, 8 are green, and 6 are yellow. We draw three balls from this urn without replacing so that at each stage (draw) any of the existing balls in the urn is equally likely to be chosen. Find the probability that the first and third draws are both green. How about both the same color.

Where do we start?

Well... visualize yourself performing this experiment as though you were the person asking these questions. Maybe start by drawing a picture of the urn on a piece of paper. Something like this:



Since we are interested in a probability here, we should first settle on a sample space Ω. Since the event of interest is green on first and third draw, it should be clear that we need a sample space where we know the color ball at each draw. So treating the draws as stages we might model the sample points as ordered 3-tuples (x, y, z), where x is the color drawn first, y the color drawn second, and z the color drawn third. But now we have an issue: there are more red balls than green, so it appears that it is more likely to select red at any draw than either of the other colors. E.g., (r, r, r) is more likely than, say, (g, g, g). But, equally-likely here should mean that any of the 30 balls has the same chance of being selected first, then any of the remaining 29 balls should have the same chance of being selected second, and any of the remaining 28 balls should have the same chance of being selected third. So, let's mark the 16 red balls as $r_1, r_2, \ldots, r_{16}$, the 8 green balls as $g_1, g_2, \ldots, g_8$ and the 6 yellow balls as $y_1, y_2, \ldots, y_6$, and think of the urn as

$$U := \{r_1, r_2, \ldots, r_{16}, g_1, g_2, \ldots, g_8, y_1, y_2, \ldots, y_6\}.$$

We get around the issue now by modeling ω as (x, y, z) where $x \in U$, $y \in U - \{x\}$, and $z \in U - \{x, y\}$. The sample space of all these ω will be equally-likely. *What's* $|\Omega|$*?*

This fits the paradigm of the basic counting rule: the answer is $|\Omega| = 30 \times 29 \times 28$.

Now we ask: with this model of Ω, *is it easy to count the number of sample points in the event* A *that the first and third draw are both green?*

Try to envision what sample points look like in A. Also, think about the process that leads to an $\omega \in A$: namely, pick a green first, pick anything remaining second, and pick a green third. Formally, by setting $G = \{g_1, g_2, \ldots, g_8\}$,

$$A = \Big\{(x, y, z) : x \in G, y \in U - \{x\}, z \in G - \{x, y\}\Big\}.$$

Does this fit the paradigm of the basic counting rule? Not quite. It's very much like the election results problem. The number of green marbles that remain in the urn at stage 3 depends on the color of the ball chosen at stage 2. After one of the 8 green balls is chosen at stage 1, there are 29 remaining balls (7 are green and 22 are not green). If one of the 7 greens is selected at stage 2, then there are 6 greens that can be selected at stage 3. If one of the 22 non-greens is chosen at stage 2, then there are 7 greens that can be selected at stage 3. Thus,

$$|A| = 8 \cdot 7 \cdot 6 + 8 \cdot 22 \cdot 7,$$

and our probability is

$$P(A) = \frac{8 \cdot 7 \cdot 6 + 8 \cdot 22 \cdot 7}{30 \cdot 29 \cdot 28}.$$

There's now something worth noting in this calcuation: $8 \cdot 7 \cdot 6 + 8 \cdot 22 \cdot 7 = 8 \cdot 7 \cdot (6 + 22) = 8 \cdot 7 \cdot 28$ so that the probability above becomes

$$P(A) = \frac{8 \cdot 7 \cdot 28}{30 \cdot 29 \cdot 28} = \frac{8 \cdot 7}{30 \cdot 29},$$

which counts the number of sample points where the first and *second* are green, i.e., there are the same number of sample points in Ω where the first and second are green

as there are sample points where the first and third are green. This peculiarity will be discussed in a bit (page 19) – it is called ***exchangeability***. But, with this it follows that the probability the first and third are green is the same as the probability the first and second are green which is:

$$\frac{8 \cdot 7}{30 \cdot 29}.$$

I'll leave you to think about why the event C that the first and third are the same color has probability

$$\begin{aligned} P(C) &= \frac{(16 \cdot 15 \cdot 14 + 16 \cdot 14 \cdot 15) + (8 \cdot 7 \cdot 6 + 8 \cdot 22 \cdot 7) + (6 \cdot 5 \cdot 4 + 6 \cdot 24 \cdot 5)}{30 \cdot 29 \cdot 28} \\ &= \frac{(16 \cdot 15) + (8 \cdot 7) + (6 \cdot 5)}{30 \cdot 29}. \end{aligned}$$

The second equality here emphasizes that the probability is really the same as the probability the first two balls drawn are the same color.

Important special cases of the basic counting rule.

1. Sampling with replacement.
From n distinct objects, how many sequences of length $k \geq 1$ can be made where we allow repetition of objects.

There are $\underbrace{n \times n \times \cdots \times n}_{k \text{ times}} = n^k$ possible selections.

Examples.

• If you toss a coin k times, how many sequences are possible?

Solution: Here $n = 2$, $\{h, t\}$, therefore, there are 2^k possible sequences.

• If you roll a 6-sided die k times*, how many sequences are possible?

Solution: Here $n = 6$, $\{1, 2, 3, 4, 5, 6\}$, therefore, there are 6^k possible die rolls.
* We keep track of the order of the rolls or, equivalently, we think of the k dice as different colors and observe the up-face as well as which color it appeared on. For example, if $k = 2$, then we can think that the first roll as the result on the red die and the second roll as the result on the green die. In this way, when we roll these (distinguishable) multicolored dice *simultaneously*, the sequence $(6, 1)$ means 6 on red, and 1 on green and clearly this is a different outcome than $(1, 6)$. Later on we will discuss counting in the situation where we roll several *indistinguishable* dice simultaneously. In this case we will be counting ***multisets*** and we use an idea called ***stars and bars*** type counting.

• How many ways can a person answer a 9-question True/False exam (assuming they answer each question)?

Solution: Here, $n = 2$, $\{T, F\}$, moreover, the stages are 'answer to question 1', 'answer to question 2',...,'answer to question 9', Therefore, we can think of the sample points as ordered 9-tuples, where each entry is either True or False. So, 2^9 possible answer sheets.

(*continued*) What if leaving questions blank is allowed?

Solution: Now, $n = 3$, $\{T, F, \text{blank}\}$, and therefore, there are 3^9 possible answer sheets.

• 15 employees sign up for exactly one of 6 possible jobs. How many different sign-ups are possible?

Before providing the solution, think about actually performing this experiment. What would an outcome of this experiment look like? We are trying to count *sign-ups*. What's a sign-up? Well...we model it! I'm thinking that the 15 employees are lined up, and, when they reach the front of the line, they declare their job choice. So, from this point of view, I model the sample point as an ordered 15-tuple where each entry is one of 6 jobs. Thus, $n = 6$ and $k = 15$. There are 6^{15} sign-ups.

2. The number of ways to order n distinct objects.

$$n! := n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1.$$

By convention we set $0! \equiv 1$. The expression $n!$ is pronounced n ***factorial***.

3. Sampling *without* replacement.
From n distinct objects, how many sequences of length $1 \leq k \leq n$ can be made where there are no repeated entries?

There are

$$\underbrace{n \times (n-1) \times (n-2) \times \cdots \times (n-(k-1))}_{k \text{ factors}} =: (n)_k = {}_nP_k = \frac{n!}{(n-k)!}$$

sequences (orderings, arrangements).

When $k = 0$, by convention, we would have only the empty sequence, so $(n)_0 \equiv 1$. But, there are *no* sequences of length greater than n, therefore, we define $(n)_k = 0$ when $k > n$, and $(n)_n = n!$. The expression $(n)_k$ is pronounced n ***falling factorial*** k.

Examples.

• There are 4 slips of paper (numbered 1,2,3,4) in a hat. The experiment is to draw all slips of paper one at a time (without replacing) noting the number on each.

Solution: There are 4! (orderings) ways to pull out all the slips of paper. Since $4! = 24$ is rather small we can list out all $\omega \in \Omega$. See experiment 3.1 on page 7.

(*continued*) What if we draw only 2 slips of paper instead?

Solution: $n = 4$ distinct objects, forming sequences of length $k = 2$. Therefore, $(4)_2 = 4 \cdot 3 = 12$. Again, this number is rather small and we list out the 12 possibilities in experiment 3.2 on page 7.

• In a standard deck of 52 cards: 13 ranks $(2, 3, 4, 5, 6, 7, 8, 9, 10$, Jack, Queen, King, Ace) in each of 4 suits $(\clubsuit, \diamondsuit, \heartsuit, \spadesuit)$. How many ways can these cards be ordered?

Solution: 52!. This is a *very* large number! For fun: type 'How large is 52!' in a Google search engine. Although it would be impossible list out all 52! ω's, it shouldn't be difficult to understand the process that leads to them.

• 30 balls: 16 red $\{r_1, r_2, \ldots, r_{16}\}$, 8 green $\{g_1, g_2, \ldots, g_8\}$, and 6 yellow $\{y_1, y_2, \ldots, y_6\}$. We draw *all* the balls one at a time without replacing. How many orderings can we observe?

Solution: There are 30! ordering of these 30 distinct balls.

(*continued.*) If, instead, we draw only 3 balls (without replacing) how many sequences can we observe?

Solution: There are $(30)_3 = 30 \cdot 29 \cdot 28$ orders of 3 balls we can observe drawn from these 30.

(*continued*) Now, a more ***advanced question***: How many orderings have all the balls of like color together? For example,

$$\underbrace{g_3, g_7, \ldots, g_5}_{\text{all green}}, \underbrace{r_6, r_{14}, \ldots, r_2}_{\text{all red}}, \underbrace{y_5, y_1, \ldots, y_2}_{\text{all yellow}}$$

is one such ordering.

Solution: Clearly, 30! *is way too much* since a sequence among the 30! can have the colors alternating, for instance. So, let's think about the process that leads to a sample point in this event. We want all the balls of the same color grouped next to each other. In the sample point demonstrated above we have green, followed by red, followed by yellow *and* within each color we have an ordering of the balls of that color. But, it seems clear that in the above sample point, by taking the exact same ordering within each color but simply moving the blocks of colors around, we would get a different sequence. So, we look at a fixed ordering of colors, say like the one above: green, red, yellow. And ask: for this ordering of colors how many orderings of the balls are there. The process to order the balls is 3 stages. In stage 1 (green) we select an order of green balls – there are 8! possible. For each such ordering of the green balls, stage 2 (red) has us select an order of the red balls – there are 16! such; and, for each ordering or green and red balls, stage 3 (yellow) has us select an ordering for the yellow balls – there are 6! such ordering. The basic counting rule tells us that there are 16!8!6! orderings of the balls when the color sequence is green, red, yellow. But now, every ordering of the 3 colors yields the same number of ordering of the balls. And, since there are 3! ways to order the 3 colors and 16!8!6! ways to order the balls of like color amongst themselves, it follows there are 3!16!8!6! ordering of the balls that have all balls of like color together.

Example and discussion. (a first illustration of the idea of exchangeability)
Suppose we have 5 chips numbered 1 through 5 in a hat, and we plan to select 3 of them without replacement one at a time. All the possible sample points are listed in the following figure (we assume they are all equally likely):

123	213	312	412	512
124	214	314	413	513
125	215	315	415	514
132	231	321	421	521
134	234	324	423	523
135	235	325	425	524
142	241	341	431	531
143	243	342	432	532
145	245	345	435	534
152	251	351	451	541
153	253	352	452	542
154	254	354	453	543

Figure. The sample points ω of Ω listed out.

In the above representation, the sample point 324, for instance, means chip number 3 was drawn first, chip number 2 drawn second, and chip number 4 drawn third. I ask:

What's the probability that chip number 3 is drawn first?

Intuitively, the answer is just $\frac{1}{5}$ since any of the 5 numbers is equally likely to have been chosen first; i.e., we simply view the experiment as stopping after the first draw. Alternatively, we can work with the sample space Ω and note $|\Omega| = 5 \cdot 4 \cdot 3$, while the event of interest has $1 \cdot 4 \cdot 3$ sample points, and the classical probability measure dictates the probability as $\frac{1 \cdot 4 \cdot 3}{5 \cdot 4 \cdot 3} = \frac{1}{5}$.

Now, I ask:

What's the probability that chip number 3 is drawn *third*?

Intuitively, the answer should also be $\frac{1}{5}$ because before the experiment is actually performed the number 3 should be equally likely to occur in any of the 3 positions; in fact, each of the 5 numbers should be equally likely to occur in any of the 3 positions.

Interestingly, there is another way to view this. What would happen to our original sample space if, in each sample point above, we simply *exchange* the first and third entry. Then, in this *exchanged* sample space, the first entry of a sample point can be viewed as the chip number that was drawn third (and, the third entry as the chip number drawn first). But, here's the amazing thing: the exchanged sample space and the original sample space are *exactly the same set!* The next figure has the sample points in the exchanged Ω listed out ω for ω (convince yourself the two lists are the same):

321	312	213	214	215
421	412	413	314	315
521	512	513	514	415
231	132	123	124	125
431	432	423	324	325
531	532	523	524	425
241	142	143	134	135
341	342	243	234	235
541	542	543	534	435
251	152	153	154	145
351	352	253	254	245
451	452	453	354	345

Figure. The sample points of the exchanged Ω listed out.

Now, the point is this: Because these sample spaces are the same, we can view the original sample space – just the way it is shown at the start of this example – as having first entry showing what was drawn third. And, now, it should be clear that there are just as many sample points in the event that chip number 3 drawn third as there are sample points in the event that chip number 3 drawn first, i.e., these two events have the same cardinality. Moreover, since all sample points are equally likely, the probabilities of these events must be the same as well!

As a thought exercise, it should be easy to see that had this experiment been to draw chips *with replacement*, then the same is true; namely, if the experiment had been to draw 3 chips with replacement after each draw, then the probability the first drawn is 3 would be the same as the probability the third drawn is 3. But, this is not really all that interesting.

Remark.
The ideas just presented remain true when we have n chips numbered 1 through n and we draw k chips without replacement, where $1 \leq k \leq n$. But, also, the ideas can be generalized significantly. We have a sample space Ω of all possible sequences of length k where entries are any of n objects without repetition. Instead of exchanging the first and last entry of each sequence in Ω as we did in the example, we can, in fact, apply any fixed permutation to each sample point in Ω and recover the *same* sample space. For instance, one choice is the permutation (mapping) that reverses the order of the each $\omega \in \Omega$:

$$\pi : (x_1, x_2, x_3, x_4, x_5, x_6) \mapsto (x_6, x_5, x_4, x_3, x_2, x_1).$$

Another possibility is the permutation that simultaneously swaps the first and third entry *and* the second and fifth entry:

$$\pi : (x_1, x_2, x_3, x_4, x_5, x_6) \mapsto (x_3, x_5, x_1, x_4, x_2, x_6).$$

In either of these instances the "permuted" sample space and the original sample space will be identical! If you've had discrete math or any exposure to proof writing, you should try to prove this for yourself.

Example.
We have 8 blue marbles $(b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8)$ and 5 red marbles $(r_1, r_2, r_3, r_4, r_5)$. We draw 6 marbles without replacement, where, on each draw, a marble is selected equally likely among the remaining marbles. Compute the probability that the last marble you draw is red. Compute the probability that the third marble is blue and the fifth marble is red.

Under the conditions of this experiment our sample space Ω is finite and equally likely, $|\Omega| = (13)_6$ by the basic counting rule.

Let A be the event where the sixth marble drawn is red. Again, intuitively, $P(A) = \frac{5}{13}$ since a red marble is equally likely to appear in any position and there are 5 red out of 13 total marbles; but, given the previous remark, we can apply the permutation that reverses the order of every $\omega \in \Omega$, get a sample space that is the same as the original sample space so that $|A|$ is the same size as the event that the first marble drawn is red. Consequently, $P(A) = \frac{5}{13}$.

Let B be the event where the third is blue and the fifth is red. To compute $P(B)$ we recognize that applying the permutation that simultaneously swaps positions 3 and 1 and swaps positions 5 and 2 to every $\omega \in \Omega$ gives the original sample space back again. But then, $|B|$ is the same size as the event that has the first marble drawn being blue and the second marble drawn red, and the cardinality of this event is, by the basic counting rule,

$$8 \cdot 5 \cdot 11 \cdot 10 \cdot 9 \cdot 8.$$

So, $P(B) = \frac{8 \cdot 5 \cdot 11 \cdot 10 \cdot 9 \cdot 8}{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9 \cdot 8} = \frac{8 \cdot 5}{13 \cdot 12}$, which, not surprisingly, is the same as the probability that the first marble is blue and the second is red by stopping the experiment after the second draw.

Anagrams.

An ***anagram*** is an ordering of the "letters" in a "word".

Example.
How many anagrams of the word **MATH** are possible? List them all.

Solution: This is an old question. In fact, this is exactly the same as experiment 3.1 on page 7: since there are 4 distinct letters in this word, the number of orderings should be just $4! = 24$. So, if we do things correctly we should have 24 orderings. Here goes:

$MATH$	$MAHT$	$MTAH$	$MTHA$	$MHAT$	$MHTA$
$AMTH$	$AMHT$	$ATMH$	$ATHM$	$AHMT$	$AHTM$
$TMAH$	$TMHA$	$TAMH$	$TAHM$	$THMA$	$THAM$
$HMAT$	$HMTA$	$HTMA$	$HTAM$	$HAMT$	$HATM$

The reason the last example was "easy" was because the 4 letters in MATH are *distinct* so that we are able to recognize the counting as just sampling all $k = n = 4$ letters without replacement. The same reasoning explains why there are 7! anagrams of the word DISPARE.

We now investigate what happens when the letters are not all distinct.

Example and discussion.
How many anagrams of the word **PEEP** are possible?

An immediate issue you may see is that when we swapped the first and last letter in MATH we obtained a distinct ordering HATM, but if we swap the first and last letter in the word PEEP we obtain the *same* word. So, in this problem, it appears 4! would be over-counting because 4! would be the number of anagrams had the letters in the word PEEP been all distinct, which they are not.

To count the actual number of anagrams we first artificially make these 4 letters in PEEP distinct, say, by putting subscripts on the letters: $\mathrm{P_1E_1E_2P_2}$, then there *are* 4! anagrams of this "word". Here are 24 anagrams grouped by common spelling:

$P_1E_1E_2P_2$	$P_1E_1P_2E_2$	$P_1P_2E_1E_2$	$E_1P_1E_2P_2$	$E_1P_1P_2E_2$	$E_1E_2P_1P_2$
$P_1E_2E_1P_2$	$P_1E_2P_2E_1$	$P_1P_2E_2E_1$	$E_2P_1E_1P_2$	$E_2P_1P_2E_1$	$E_2E_1P_1P_2$
$P_2E_1E_2P_1$	$P_2E_1P_1E_2$	$P_2P_1E_1E_2$	$E_1P_2E_2P_1$	$E_1P_2P_1E_2$	$E_1E_2P_2P_1$
$\underbrace{P_2E_2E_1P_1}$	$\underbrace{P_2E_2P_1E_1}$	$\underbrace{P_2P_1E_2E_1}$	$\underbrace{E_2P_2E_1P_1}$	$\underbrace{E_2P_2P_1E_1}$	$\underbrace{E_2E_1P_2P_1}$
all words spell PEEP	all words spell PEPE	all words spell PPEE	all words spell EPEP	all words spell EPPE	all words spell EEPP

When grouped this way it is clear that there are always 4 words in each column that all spell the same word. I.e., in each column the 2 P's and the 2 E's are in the same relative positions, and therefore, if we remove the subscripts then all words in a column are the same. Therefore, the number of columns, i.e., $\frac{4!}{4} = 6$, *is* the number of distinct spellings of the word PEEP (without subscripts). So, there are 6 anagrams of PEEP.

One way to have seen that there are always going to be 4 orderings of the letters $\mathrm{P_1}$, $\mathrm{E_1}$, $\mathrm{E_2}$, $\mathrm{P_2}$ that all spell the same word is as follows. Think of the process that keeps the 2 P's ($\mathrm{P_1}$ and $\mathrm{P_2}$) and the 2 E's ($\mathrm{E_1}$ and $\mathrm{E_2}$) in the same relative positions. For instance, fix the 2 positions for the P's (the remaining 2 positions will have to be for the E's) for a word. The process is 2 stages: stage 1 has us put $\mathrm{P_1}$ and $\mathrm{P_2}$ into the 2 fixed positions (there are 2! ways to do this), and, for each way that stage 1 is completed, stage 2 has us place to 2 E's (also 2! ways). Thus, there are $2!2! = 4$ orderings of $\mathrm{P_1E_1E_2P_2}$ that have the same spelling.

The last discussion can lead us to provide a proof for...

Counting the number of anagrams.
If an n-letter word is comprised of $k \geq 1$ types of *indistinguishable* letters of which there are n_i of type i $(i = 1, \ldots, k)$, where $n_1 + n_2 + \cdots + n_k = n$, then there are

$$\frac{n!}{n_1! n_2! \cdots n_k!}$$

anagrams.

Examples.

• How many anagrams of INDEPENDENCE are there?

SOLUTION:
This is an $n = 12$ letter word.
There are $k = 6$ types of letters (I,N,D,E,P,C),
$n_1 = 1$ I, $n_2 = 3$ N's, $n_3 = 2$ D's, $n_4 = 4$ E's, $n_5 = 1$ P, and $n_6 = 1$ C.
Therefore, there are $\dfrac{12!}{1!3!2!4!1!1!} = \dfrac{12!}{4!3!2!}$ anagrams.

(*continued*) How many anagrams of INDEPENDENCE begin NE (in this order)?

SOLUTION:
Think of the process that leads to such an anagram. Each such anagram will begin NE followed by the remaining letters in some order. If you try to visualize the list of all anagrams that look like this, then it should not be hard to see that the number of them will be the same as if we just *erase* the NE at the beginning of each. We'd be now left with all the anagrams of a 10-letter word with 1 I, 2 N's, 2 D's, 3 E's, 1 P, and 1 C. Therefore, there are $\dfrac{10!}{1!2!2!3!1!1!} = \dfrac{10!}{3!2!2!}$ anagrams.

(*continued*) How many anagrams will have all the vowels contiguous (i.e., grouped together)?

SOLUTION:
Here's a slick way: since all the vowels will be grouped together, let's temporarily create the super-letter V, i.e., V represents IEEEE. Now, let's remove the vowels from INDEPENDENCE (leaves us with NDPNDNC) and put the *one* super-letter V in the vowels absence: VNDPNDNC, an 8-letter word with 1 V, 3 N's 2 D's, 1 P, and 1 C, and there are $\dfrac{8!}{1!3!2!1!1!} = \dfrac{8!}{3!2!}$ such anagrams. Now, in these anagrams, if we replace the V with IEEEE in its place, we'd have all the anagrams of INDEPENDENCE with the vowels grouped together in the order IEEEE. But, of course, these vowels can be re-ordered in $\dfrac{5!}{1!4!} = 5$ ways. So, the number of anagrams is $5 \cdot \dfrac{8!}{3!2!}$.

• How many anagrams of 1111000 are there?

SOLUTION:

A 7-letter word with 4 1's and 3 0's. Thus, $\frac{7!}{4!3!}$ anagrams.

• We toss a coin 7 times and observe the sequence of heads and tails we get. How many sequences have exactly 4 heads?

SOLUTION:

This is just the previous question (do you see?). Answer: $\frac{7!}{4!3!}$.

• We have 30 balls: 16 red, 8 green, and 6 yellow. We line up all 30 balls left to right. How many color sequences can we observe?

SOLUTION:

Notice the subtle difference in this problem compared with what we were dealing with on page 13. In this problem, if we swap two balls of the same color we'll get the same color sequence. We are now treating balls of the same color as *indistinguishable*. For instance, the color sequence

$$GGGGGGGGRRRRRRRRRRRRRRRRYYYYYY$$

has all the greens appearing first, followed by all the reds, followed by all the yellows, and, if we permute the G's amongst their relative positions, we end up with the same color sequence. We think of the color sequence as a word of length 30 with 16 R's, 8 G's, and 6 Y's, then every color sequence corresponds to an anagram of this word. Therefore, there are

$$\frac{30!}{16!8!6!}$$

color sequences.

• 4 Germans, 4 Danes, 3 Americans, and 2 Canadians enter a road race. Everyone finishes and there are no ties. How many ways can these people finish the race if we record only their nationalities?

SOLUTION:

There are 13! finishes possible if the 13 people were treated distinct. However, in this problem since we only record their racer's nationality, all Germans are indistinguishable, all Danes are indistinguishable, etc. We are trying to count the number of possible orderings of these nationalities, i.e., which nationalities finish 1st, 2nd, 3rd, and so on. But, this is just the number of anagrams of the word GGGGDDDDAAACC. There are

$$\frac{13!}{4!4!3!2!}$$

possible finishes.

EXERCISES.

1. A 7-sided die is rolled 14 times. How many sequences of outcomes are there?

2. Noelle has a basket of 8 toys for her dog. She pulls a toy for the dog to play with, then takes that toy away from the dog and puts it back into the basket. She then grabs one of the those toys again uniformly at random. She repeats this a total of 6 times. In how many ways can Noelle draw the toys?

3. Feller's *An Introduction to Probability Theory and Its Applications, Volume 2* has 670 pages. Dr. Torcaso opens the book, flips to a random page, and then closes the book. This is repeated 10 times. How many different sequences of pages can Dr. Torcaso obtain?

4. How many subsets of $\{1, 2, \ldots, n\}$ exclude the subset $\{1, 2, \ldots, k\}$, where $1 \leq k \leq n$?

5. There are 8 soccer players on a Hall of Fame candidate list. You can select *any* number of the people as your choice(s) to enter the Hall of Fame as you like (zero is possible as well as all 8). How many different selections are there?

6. Paul has to go to the Bradford laundry room. There are 15 washers. He does his 5 loads of laundry one-at-a-time, and he can select any of the 15 washers each time. Find the number of ways Paul can do his 5 loads of laundry.

7. Adam is playing poker. On each turn, since he is aggressive and doesn't know how to fold, he either checks or raises. There are 6 bets in a given hand. Find the amount of sequences of actions that Adam can perform in one hand.

8. There are 20 sectors on a standard dartboard. Six people throw darts at random at this board. If we record what sector each person landed in, how many recordings are possible?

9. Suppose a sandwich is
choice of bread (choose one: 1=rye, 2=wheat, 3=white, 4=kaiser roll)
choice of protein (choose one: 1=roast beef, 2=turkey, 3=tuna salad)
choice of cheese (choose one: 1=American, 2=Swiss, 3=no cheese)
choice of lettuce (choose one: 1=yes, 2=no)
choice of tomato (choose one: 1=yes, 2=no)
(a) How many sandwiches on rye are possible?
(b) How many roast beef sandwiches have cheese?
(c) If a friend says they will buy you a sandwich (so that all possible sandwiches are equally likely), what's the probability you get a roast beef sandwich?
(d) (separate question) What's the probability your friend put's cheese and lettuce on your sandwich?
(e)* If a sandwich is allowed to not have a protein (for example, two slices of bread with nothing in-between can be considered a sandwich now), how many sandwiches are possible now?
(f)** If a sandwich is now allowed to have more than one type of protein (as well as no protein) and is allowed to have more than one type of cheese (as well as no cheese), how many sandwiches are possible now?
(g) The deli worker refuses to put mustard on Tuna Salad. How many sandwiches are possible? Try to count this in two ways.

10. A manager has 165 players from which they are trying to fill a roster of 11 different positions. How many rosters are possible?

11. A basketball manager has 8 players $6'$ in height or taller and 6 players under $6'$ tall on the bench. The manager wants to create a starting roster (5 different positions) consisting of a center, a power forward, a small forward, a shooting guard, and a point guard.
(a) If there are no restrictions on how the manager can make the starting team of 5 players, how many starting rosters are possible?
(b) If only players $6'$ or taller can be rostered for the center and two forward positions, and only players under $6'$ can be rostered for the guard positions, how many starting rosters are possible now?
(c) (separate question) If only players 6' or taller can be rostered for the center and two forward positions (and no height restriction on guards), how many starting rosters are possible?

12. There are 8 horses in a race at Pimlico. The first horse to finish is ranked 1, the second horse to finish is ranked 2, and so on. All horses finish the race, there are no ties.
(a) How many rankings are possible?
(b) The rank 1, 2, and 3 positions are sometimes called the *win, place,* and *show* positions, respectively. How many win, place, show results are possible?

13. A child is putting away 12 different Lego blocks. In how many different orders can they be put into the container?

14. We have a standard deck of 52 cards well-shuffled. We turn over the top 5 cards one at a time.
(a) How many arrangements are possible?
(b) Find the probability all cards are red (the diamond $\diamondsuit$ and heart $\heartsuit$ suits are red, other suits are black).
(c) Find the probability colors alternate.
(d) (separate question) Suppose when we turn over the 5 cards we replace each card we turn over into the deck and re-shuffle before drawing the next, then how many possible arrangements are there now?

15. Suppose k and n are positive integers with $1 \leq k \leq n$. How many orderings of the n integers 1 through n have the first k entries from the set 1 through k?

16. 10 people sit at a round table. How many distinct arrangements of seats are there? Any arrangements that can be obtained by rotating the people at the table around but not changing the order of the seats are considered identical. Also, if all such seating arrangements were equally likely, what's the probability Fred is seated next to Carrie?

17. A lottery card consists of 6 distinct numbers from 1 to 90 inclusive. If the order is relevant in determining a winner, how many different lottery cards are there?

18. Gary creates a workout plan at a gym. There are 28 different machines, and the order in which he selects machines influences his workout. How many different ways can Gary create a workout consisting of 7 machines if
(a) Gary can repeat a machine at any point in the workout?
(b) Gary cannot repeat a machine in the workout?
(c) Gary can repeat a machine in the workout but just not consecutively?

19. Consider the word BOOLAHUBBOO.
(a) How many anagrams are possible?
(b) How many of these anagrams end BOOBOO?
(c) How many anagrams have all the B's grouped together?
(d) How many anagrams have all the B's grouped together and all the vowels grouped together?

20. I have 3 one dollar bills, 2 five dollar bills and 5 ten dollar bills. How many ways can I distribute these 10 bills to 10 children so that each child gets one bill? Treat the bills of equal value as indistinguishable please.

21. There are 9 people who will toss a ball around. The only rule is that you cannot toss to yourself. One of them picks up a ball, tosses to another, who tosses to another (we're allowed to toss back to the person who tossed to them), and this repeats. E.g., if a picks up the ball, tosses to b, who tosses to a, who tosses to c, we would observe the ordered 4-tuple (a, b, a, c) of *three* tosses.
(a) Let Ω be the set of all such lists when we have 9 people and 7 tosses. Compute $|\Omega|$.
(b) Assume the sample space Ω of all 7-toss possibilities are equally likely. What's the probability that f picks up the ball then tosses to a, b, or c?
(c) Give an example of a permutation which, when applied to all $\omega \in \Omega$, will *not* give the same sample space back again.
(d)* (challenging) With 9 people and 7 tosses, how many possible lists can be observed where Fred is the last to catch the ball?

Up to now, we've only considered experiments that generated sample points modeled through stages as a process, and we've found that the basic counting rule has an enormous number of applications in this regard. However, there are some experiments where the order in which stages are performed is not needed.

Here's a fairly abstract but prototypical example of what I'm talking about. Consider a set of n distinct objects. Without loss of generality we can think of the set as

$$\{x \in \mathbb{Z} : 1 \leq x \leq n\} = \{1, 2, 3, \ldots, n\}.$$

How many subsets of size k $(0 \leq k \leq n)$ are there from this set?

Combinations.
There are

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

subsets of size k from the n-element set. The symbol $\binom{n}{k}$ is pronounced n ***choose*** k and, in mathematics, it is referred to as a ***binomial coefficient***.

$\binom{n}{k}$ counts how many ways we can sample k objects without replacing, ignoring order.

Example and discussion.
How many subsets of size 3 are there from a 5-element set?

SOLUTION: There are $\binom{5}{3} = \frac{5!}{3!2!} = \frac{5 \cdot 4 \cdot (3!)}{3!2!} = \frac{5 \cdot 4}{2 \cdot 1} = 10$. Here are the 10 subsets listed out:

$$\{1,2,3\}\ \{1,2,4\}\ \{1,2,5\}\ \{1,3,4\}\ \{1,3,5\}\ \{1,4,5\}\ \{2,3,4\}\ \{2,3,5\}\ \{2,4,5\}\ \{3,4,5\}$$

Had we decided to count the subsets by first including order we would end up with the following $(5)_3 = 60$ ordered 3-tuples:

$(1,2,3)$	$(1,2,4)$	$(1,2,5)$	$(1,3,4)$	$(1,3,5)$	$(1,4,5)$	$(2,3,4)$	$(2,3,5)$	$(2,4,5)$	$(3,4,5)$
$(1,3,2)$	$(1,4,2)$	$(1,5,2)$	$(1,4,3)$	$(1,5,3)$	$(1,5,4)$	$(2,4,3)$	$(2,5,3)$	$(2,5,4)$	$(3,5,4)$
$(2,1,3)$	$(2,1,4)$	$(2,1,5)$	$(3,1,4)$	$(3,1,5)$	$(4,1,5)$	$(3,2,4)$	$(3,2,5)$	$(4,2,5)$	$(4,3,5)$
$(2,3,1)$	$(2,4,1)$	$(2,5,1)$	$(3,4,1)$	$(3,5,1)$	$(4,5,1)$	$(3,4,2)$	$(3,5,2)$	$(4,5,2)$	$(4,5,3)$
$(3,1,2)$	$(4,1,2)$	$(5,1,2)$	$(4,1,3)$	$(5,1,3)$	$(5,1,4)$	$(4,2,3)$	$(5,2,3)$	$(5,2,4)$	$(5,3,4)$
$(3,2,1)$	$(4,2,1)$	$(5,2,1)$	$(4,3,1)$	$(5,3,1)$	$(5,4,1)$	$(4,3,2)$	$(5,3,2)$	$(5,4,2)$	$(5,4,3)$

But, if we only cared about the entries and not the order they are in, then computing the number of ordered 3-tuples would be *way* over-counting! How much is $(5)_3$ over-counting by?

It seems every choice of 3 distinct elements from $\{1,2,3,4,5\}$ is repeated $3! = 6$ times in the table, when we only need *one*! Just look at the columns of the table. Every column has the same 3 entries, and we know there must be 3! of them. So $(5)_3$ is over-counting by 3! and, therefore, there are $\frac{(5)_3}{3!}$ subsets.

An extension of this argument to k-element subsets of an n-element set should now be straightforward.

Remark. (A slight digression)
Here's an interesting follow-up to the last example. Suppose our experiment is to draw 3 numbers without replacement from a hat having numbers 1,2,3,4,5. We draw so that each number is equally likely to be any of the remaining numbers in the hat.

What's the probability we draw the numbers in decreasing order?

Looking at the equally likely sample space, (we list those sample points again here:)

$(1,2,3)$	$(1,2,4)$	$(1,2,5)$	$(1,3,4)$	$(1,3,5)$	$(1,4,5)$	$(2,3,4)$	$(2,3,5)$	$(2,4,5)$	$(3,4,5)$
$(1,3,2)$	$(1,4,2)$	$(1,5,2)$	$(1,4,3)$	$(1,5,3)$	$(1,5,4)$	$(2,4,3)$	$(2,5,3)$	$(2,5,4)$	$(3,5,4)$
$(2,1,3)$	$(2,1,4)$	$(2,1,5)$	$(3,1,4)$	$(3,1,5)$	$(4,1,5)$	$(3,2,4)$	$(3,2,5)$	$(4,2,5)$	$(4,3,5)$
$(2,3,1)$	$(2,4,1)$	$(2,5,1)$	$(3,4,1)$	$(3,5,1)$	$(4,5,1)$	$(3,4,2)$	$(3,5,2)$	$(4,5,2)$	$(4,5,3)$
$(3,1,2)$	$(4,1,2)$	$(5,1,2)$	$(4,1,3)$	$(5,1,3)$	$(5,1,4)$	$(4,2,3)$	$(5,2,3)$	$(5,2,4)$	$(5,3,4)$
$\mathbf{(3,2,1)}$	$\mathbf{(4,2,1)}$	$\mathbf{(5,2,1)}$	$\mathbf{(4,3,1)}$	$\mathbf{(5,3,1)}$	$\mathbf{(5,4,1)}$	$\mathbf{(4,3,2)}$	$\mathbf{(5,3,2)}$	$\mathbf{(5,4,2)}$	$\mathbf{(5,4,3)}$

every choice of 3 numbers we could have drawn has exactly 3! orderings. Moreover, among the 3! equally likely orderings, only 1 of them will be in decreasing order (***bold-faced*** in table above). Therefore, the probability we draw the numbers in decreasing order is $\frac{1}{3!}$.

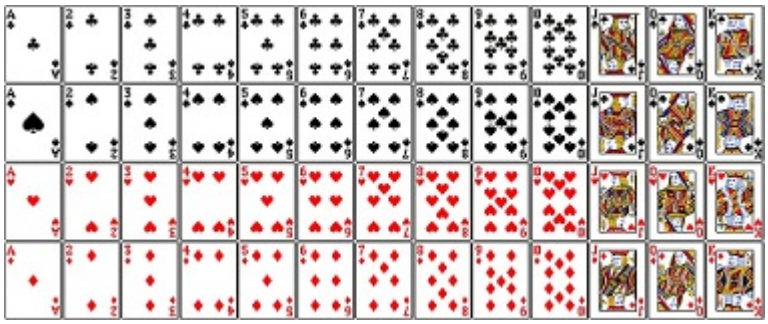
Alternatively, there are $\binom{5}{3}$ ordered 3-tuples in decreasing order out of a sample space of $(5)_3$ equally likely sample points. Thus, the probability is also $\dfrac{\binom{5}{3}}{(5)_3}$. You should verify this expression reduces to $\dfrac{1}{3!}$. It's interesting to note that this probability only depends on the size of the subset and not the number of elements we are choosing them from.

Remark.
There are *many* experiments – even ones that are modeled in stages where we may want to count sample points (or the number of ways a stage can be completed) – that ignore order, i.e., count subsets. Here are some examples.

Example.
A committee of size 3 is to be made from Amir, Betty, Cristoff, Dahlia, and Ernesto. How many committees are possible?

SOLUTION: Unlike the situation where 3 of them were being elected into distinguishable positions (president, vice-president, treasurer), here we have the 3 members serving a common committee – we can think of the positions as indistinguishable. So, we are counting subsets. There are $\binom{5}{3}$ committees.

Standard deck of 52 cards.



Figure[1]**.** 13 ranks:A,2,3,...J,Q,K; 4 suits: Club(♣), Spade(♠), Heart(♡), Diamond(♢)

A ***k-card hand*** or, simply, ***hand*** of k cards is a subset of size k from from the deck.

Example.
From a standard deck of 52 cards, how many 5-card hands are possible?

SOLUTION: There are $\binom{52}{5}$ such hands. As an integer this is 2,598,960 hands.

Remark. (Be careful when picking models to work with cards!)
A problem might tell you *loosely* that someone is dealing you *a hand* when they might really mean they are dealing you *a sequence of* 5 cards. We should really understand the questions being asked before blindly using the $\binom{52}{5}$ sample space. For example, if someone asks for the probability the first card is a King while the last card is a club, then this would require a sample space that keeps track of the order in which cards are dealt; otherwise, this would not be an event in (i.e., a subset of) the $\binom{52}{5}$ sample space. And, what the problem called a *hand* should really be thought of as an ordered 5-tuple of cards dealt without replacement. In this case, we should (and, possibly, need to) use the sample space that has $(52)_5$ equally likely sample points.

However, we may have a choice in deciding whether or not to have a sample space keep track of order. For example, consider the events

A: we are dealt 5 red cards, and

B: we are dealt 3 black and 2 red cards.

For these events, we do not need to know the order the cards were dealt to determine whether or not the event occurred. In these cases, we are justified (but not obliged) to model our sample points *ignoring* the order in which cards were dealt – and doing so will typically make the computations *much* easier! We can use the $\binom{52}{5}$ sample space Ω. This Ω will still have equally likely sample points.

[1]Source of Figure: https://www.milefoot.com/math/discrete/counting/cardfreq.htm

Example. Consider the event A from the previous remark.
What's the probability we are dealt 5 red cards?

SOLUTION: There are 26 red cards (and 26 black cards) in a standard deck. The occurrence (or nonoccurrence) of the event can be determined if we ignore the order of the deal, therefore, we can use the $\binom{52}{5}$ *equally likely* sample space. Since there are $\binom{26}{5}$ 5-card hands that are all red, the probability is $\dfrac{\binom{26}{5}}{\binom{52}{5}} = \dfrac{65780}{2598960} \approx 0.0253$.

Note: in this example, I'll concede that we could have also (easily) computed this probability as $\dfrac{(26)_5}{(52)_5} = 0.0253\ldots$ where we kept track of the order in which the red cards were dealt, and although it was just as easy, it wasn't necessary. But please read the next example.

Example. Consider the event B from the previous remark.
What's the probability we are dealt 3 black and 2 red cards?

SOLUTION: Since the event of getting 3 black and 2 red cards doesn't depend on the order in which these cards were dealt, I choose to work with the $\binom{52}{5}$ sample space. We need to count the number of subsets that have 3 black and 2 red cards. This is only a little tricky. Here's the idea. Label the 26 black cards $b_1, b_2, \ldots, b_{26}$ and the 26 red cards $r_1, r_2, \ldots, r_{26}$. Every subset of size 5 from the 52 that have 3 black and 2 red looks like $\{b_i, b_j, b_k\} \cup \{r_m, r_n\}$, where the indices i, j, k are distinct and the indices m, n are distinct, i.e., in stage 1 we select a subset of 3 black cards from 26 (there are $\binom{26}{3}$ ways to complete this) and, for each such selection of the black cards, in stage 2 we select 2 cards from the 26 red cards (which can be completed in $\binom{26}{2}$ ways). The basic counting rule says there are $\binom{26}{3} \cdot \binom{26}{2}$ 5-card hands with 3 black and 2 red. The probability is

$$\frac{\binom{26}{3} \cdot \binom{26}{2}}{\binom{52}{5}} \approx 0.325.$$

Just to foreshadow, in the last two probability computations we used what's called the ***hypergeometric distribution*** (see page 100).

What if we chose to keep track of order, i.e., work with the $(52)_5$ *sample space instead?* Now, we're counting ordered 5-tuples of 3 black and 2 red cards. Careful: the answer is not $(26)_3(26)_2$. This *under-counts* by a bit, since it only considers the case of a fixed ordering for the positions of black and red cards; for instance, *bbbrr*, or *brbrb*, etc. Each re-ordering will give $(26)_3(26)_2$ 5-tuples. So, we count the number of ways to re-order the positions, the basic counting rule gives us our answer. There are $\frac{5!}{3!2!}$ ways to re-order the positions of the black and red cards: this is just the number of anagrams of a 5-letter word having 3 B's and 2 R's. Therefore, there are $(26)_3(26)_2 \cdot \frac{5!}{3!2!}$ ways to receive 3 black and 2 red 5-tuples, and the probability is

$$\frac{(26)_3(26)_2 \cdot \dfrac{5!}{3!2!}}{(52)_5} = \frac{\dfrac{(26)_3}{3!}\dfrac{(26)_2}{2!}}{\dfrac{(52)_5}{5!}} =: \frac{\binom{26}{3}\binom{26}{2}}{\binom{52}{5}}.$$

Moral of the story: some problems are more straightforward when we can ignore order.

Continuing with cards: two 5-card poker hand examples...

Example.
What's the probability of getting a *full-house*? FYI: a ***full-house*** in poker is 3 cards of a rank and 2 cards of another (different) rank.

So, for instance, $\{A\clubsuit, A\diamondsuit, A\spadesuit, K\diamondsuit, K\heartsuit\}$ is a full-house – Aces over Kings.

SOLUTION: Again, we ignore the order in which cards were dealt since a full-house will be the same full-house if dealt in a different order. Now, try to envision a process that describes the sample points in this event. Here's a start: select 3 cards of one rank, and then, after this, select 2 cards of another rank – a two-stage procedure! This procedure would do it, except...

What rank do I select the 3 cards from? What rank do I select the 2 cards from?

In a full-house, the two ranks we select will happen to be *distinguishable*, i.e., if the two ranks selected are Ace and King, then we can distinguish between the full-house having 3 Aces and 2 Kings from the full-house that has 3 Kings and 2 Aces, i.e.,

$$\{A\clubsuit, A\diamondsuit, A\spadesuit, K\diamondsuit, K\heartsuit\} \neq \{K\clubsuit, K\diamondsuit, K\spadesuit, A\diamondsuit, A\heartsuit\}.$$

Order of selecting ranks matters here! So, when selecting the two ranks from the 13, we may declare the first to correspond to, say, the 3 card rank, and the second (chosen from the remaining 12 ranks) to correspond to the 2 card rank – another two-stage procedure. Finally, the process that leads to a full-house is a 4-stage procedure that fits the paradigm of the basic counting rule:

Stage 1: select rank for 3 cards – $\binom{13}{1} = 13$ choices. Once we've performed stage 1
Stage 2: select rank for 2 cards from remaining ranks – $\binom{12}{1} = 12$ choices. Then
Stage 3: select 3 cards from a rank – $\binom{4}{3}$ choices. Once we completed stage 3
Stage 4: select 2 cards from another rank – $\binom{4}{2}$ choices...

and, the number of full-houses is

$$13 \cdot 12 \cdot \binom{4}{3} \cdot \binom{4}{2}$$

and the probability of a full-house is

$$\frac{13 \cdot 12 \cdot \binom{4}{3} \cdot \binom{4}{2}}{\binom{52}{5}} = \frac{3744}{2598960} \approx 0.00144.$$

Thought exercise.
We employed the basic counting rule to both the 3 black/2 red example on page 31 and the full-house example above. Why was the counting "more involved" in the full-house example compared to the 3 black/2 red example?

Example.
What's the probability we are dealt a two-pair hand? FYI: ***two-pair*** in poker is 2 cards in a rank, 2 cards in another (different) rank (these are the *pair ranks*), and 1 card from the remaining cards that do not have the pair ranks.

So, for instance $\{A\heartsuit, A\spadesuit, K\diamondsuit, K\heartsuit, 10\heartsuit\}$ is a two-pair with pair ranks Aces and Kings.

SOLUTION: Since a two-pair hand is the same regardless of the order it was dealt, I choose to work with a sample space whose sample points exclude the order the cards are dealt. In this sample space, what's the process that leads to a two-pair?

Following what we did in the full-house example... Here's a 3-stage process:
In stage 1, select two cards of a rank; once this is done, in stage 2, we can select two cards of a different rank than in stage 1; and, once this is done, in stage 3, we can select 1 card from the remaining cards that do not have the pair ranks. But, again,

What rank do I select in stage 1? What rank do I select in stage 2?

In two-pair, unlike a full-house, the ranks we choose are *indistinguishable*! For example, the procedure that selected $\{A\heartsuit, A\spadesuit\}$ in stage 1, then selected $\{K\diamondsuit, K\heartsuit\}$ in stage 2, and then selected $\{10\heartsuit\}$ in stage 3 cannot be distinguished from the procedure that selected $\{K\diamondsuit, K\heartsuit\}$ in stage 1, then selected $\{A\heartsuit, A\spadesuit\}$ in stage 2, and then selected $\{10\heartsuit\}$ in stage 3. These are actually the *same* hand.

So, we arrive at the following 4-stage process:
Stage 1: select a *subset* of 2 ranks – $\binom{13}{2}$ choices. Then
Stage 2: select 2 cards from one of the ranks – $\binom{4}{2}$ choices. Then
Stage 3: select 2 cards from the other rank – $\binom{4}{2}$ choices. Then
Stage 4: select 1 card from the $52 - 8 = 44$ cards that do not have the pair ranks – 44 choices.
The basic counting principle says there are

$$\binom{13}{2} \cdot \binom{4}{2} \cdot \binom{4}{2} \cdot 44$$

two-pair hands, and the probability is

$$\frac{\binom{13}{2} \cdot \binom{4}{2} \cdot \binom{4}{2} \cdot 44}{\binom{52}{5}} \approx 0.0475.$$

Exercise for students.
Try re-doing the full-house and two-pair examples by working with the sample space that keeps track of order. You might just convince yourself that not keeping track of order is relatively *much simpler* and we are much less likely to make arithmetic and reasoning mistakes.

Now for a deeper example that combines some older counting ideas, and a *new* idea...

Example and discussion.
Consider the 30 balls examples from pages 13 and 24. We have 30 balls comprised of 16 red, 8 green, and 6 yellow balls. We randomly line up all the balls left to right in such a way that all line-ups are equally likely. Compute the probability that there are no two yellow balls next to each other.

Since this is a probability question we should fix a sample space. It seems natural to take the sample space Ω of all possible color sequences. We learned that $|\Omega| = \frac{30!}{16!8!6!}$ (see page 24). Please think about why Ω is equally likely.

Since the sample space of all color sequences is equally likely, we will need to count the number of color sequences that have no two yellow balls adjacent. We now think about the process that leads to such a color sequence. Here's an idea: there's no restriction on the red and green balls here, so we imagine lining up just these 24 balls (in any fashion) with enough "space" between them to potentially fit yellow balls anywhere within the sequence as well as the start and the end of the sequence. Here's a picture of what I'm talking about:

$$\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_G\_G\_G\_G\_G\_G\_G\_G\_$$

The picture has $24 + 1 = 25$ blanks that correspond to the potential places we can put yellow balls. We want to place yellow balls into the blanks in a way where there is *at most one* yellow in a blank (because, if we put two or more yellows in a blank we'd have adjacent yellows). Therefore, since there are 25 blanks, i.e., (distinct potential positions for the yellows), we just need to select any 6 of them to put our *indistinguishable* yellows into. This can be done in $\binom{25}{6}$ ways.

And, now we can count how many colors sequences have no two yellows adjacent via the basic counting rule:

Stage 1: select 6 out the 25 possible positions to put the letter Y, $\binom{25}{6}$ ways. Then

Stage 2: select an anagram of the word containing the R's and G's, $\frac{24!}{16!8!}$ ways.

Therefore, there are

$$\binom{25}{6} \cdot \frac{24!}{16!8!}$$

color sequences with no two yellows adjacent, and the probability is

$$\frac{\binom{25}{6} \cdot \frac{24!}{16!8!}}{\frac{30!}{16!8!6!}} = \frac{\binom{25}{6}}{\binom{30}{6}} \approx 0.298.$$

Remark. (Cute finding)
Notice the first equality in the probability computation *suggests* the intuitive answer to this question: The Y's can occupy $\binom{30}{6}$ possible places in the color sequence equally likely. $\binom{25}{6}$ of these colors sequences have no two Y's adjacent when ignoring the R's and G's. The ratio is our answer!

Example.
Roll a fair 6-sided die 5 times. Compute the probability of throwing a full-house.

SOLUTION: We'll take Ω to be set of all ordered 5-tuples of the numbers 1,2,3,4,5,6 with replacement. There are $|\Omega| = 6^5$ equally likely sample points. We are now going to count the number of full-houses – sample points like

$$(1,3,3,1,3), \quad (4,4,4,2,2), \quad (6,5,6,5,5), \text{ etc.}$$

In this example we are choosing to keep track of order unlike the similar situation of a hand dealt from a deck of cards. So, we should now think about the process that leads to a full-house in the current situation. Clearly, even with dice rolls, a full-house will have 3 of one rank and 2 of another (different) rank. So, similar to the card example,

Stage 1: select a rank for the triple, $\binom{6}{1} = 6$ ways to complete. Once this is done,
Stage 2: select a rank from the remaining 5 ranks for the double, $\binom{5}{1} = 5$ ways. Then
Stage 3: select the 3 positions from the 5 possible for our triple, $\binom{5}{3} = 10$ ways. Then
Stage 4: select 2 of the remaining 2 positions for our double, $\binom{2}{2} = 1$ way.

The basic counting rule says there's $\binom{6}{1} \cdot \binom{5}{1} \cdot \binom{5}{3} \cdot \binom{2}{2}$ possible full-houses. Therefore, the probability of a full-house (with 5 dice) is

$$\frac{\binom{6}{1} \cdot \binom{5}{1} \cdot \binom{5}{3} \cdot \binom{2}{2}}{6^5} = \frac{6 \cdot 5 \cdot 10 \cdot 1}{7776} = \frac{300}{7776} \approx 0.03858.$$

The reason this calculation was essentially different from the card example is that this experiment is sampling *with* replacement, whereas the card example is sampling *without* replacement.

In the last example I'll mention that had we ignored the order of the dice rolls then not all throws will be equally likely. We'll discuss this more in the Stars-and-bars counting section on page 48.

the binomial theorem.
For integer $n \geq 1$ and for any real constants a and b,

$$\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a+b)^n.$$

Notice the symmetry in the results: since $(a+b)^n = (b+a)^n$, we also have

$$\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k = (a+b)^n.$$

Taking $a = b = 1$ in the binomial theorem gives us the following:

Corollary to the binomial theorem.

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

Remark. (**Important interpretation of the corollary.**)
The total number of subsets of an n-element set is 2^n:

$$\sum_{k=0}^{n} \binom{n}{k} = \underbrace{\binom{n}{0}}_{\substack{\#\text{size } 0 \\ \text{subsets}}} + \underbrace{\binom{n}{1}}_{\substack{\#\text{size } 1 \\ \text{subsets}}} + \cdots + \underbrace{\binom{n}{n-1}}_{\substack{\#\text{size } n-1 \\ \text{subsets}}} + \underbrace{\binom{n}{n}}_{\substack{\#\text{size } n \\ \text{subsets}}} = 2^n,$$

and, since every subset of an n-element set will have exactly one of these sizes, the expression above represents the total number of subsets.

For example, if $A = \{1, 2, 3\}$ (here, of course, $|A| = 3$), then according to the corollary there must be $2^3 = 8$ subsets of A. Here they are:

$$\varnothing, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}.$$

There is $\binom{3}{0} = 1$ subset with 0 members; namely, the null set $\varnothing$.
There are $\binom{3}{1} = 3$ subsets with 1 member; namely, $\{1\}, \{2\}, \{3\}$.
There are $\binom{3}{2} = 3$ subsets with 2 members; namely, $\{1, 2\}, \{1, 3\}, \{2, 3\}$.
There is $\binom{3}{3} = 1$ subset with 3 members; namely, A itself: $\{1, 2, 3\}$,
and, of course, $1 + 3 + 3 + 1 = 8 = 2^3$.

Example.
Suppose A is a finite set with $|A| = 7$, and select a subset B ***uniformly at random***, which means all subsets of A are equally likely to be chosen. Compute the probability that $|B| = 4$.

SOLUTION: There are 2^7 possible subsets to be selected, each equally likely. The event that a subset is size 4 has cardinality $\binom{7}{4}$. Therefore, the probability is

$$\frac{\binom{7}{4}}{2^7} = \frac{35}{128} \approx 0.2734.$$

An alternate solution to this problem using older ideas goes like this:

List the 7 members of A, and flip a fair coin $n = 7$ times. If the ith is a head, put the ith member of A in the subset; else don't. Then, each sequence having exactly 4 heads corresponds to a distinct 4-element subset of A. From the example on page 24 there are $\dfrac{7!}{4!3!}$ sequences having exactly 4 heads. Moreover, there are 2^7 possible sequences which are equally likely because the coin is fair. Thus, we arrive at the same probability:

$$\frac{\dfrac{7!}{4!3!}}{2^7}.$$

This last calculation is related to the ***binomial distribution*** (see page 101) which we discuss at length in the discrete random variables section.

Example.
Suppose we have n people. We define a *club* to be any subset of these people, where one of them is identified as the *leader* along with the remaining members (possibly empty). How many clubs are possible?

Just to clarify, a club must be a subset of size *at least* 1 since a subset of size 0 would have no leader. A club can be a subset of size 1: a leader with no other members. Finally, a subset with the same members having different leaders identified are considered different clubs. So, for instance, if we have $n = 4$ people: a, b, c, d, then $\{a, b, c\}$ with a as leader and $\{a, b, c\}$ with b as leader are two *different* clubs.

SOLUTION: Here's one solution. We, again, consider a process to get a *club* as defined above. One process is as follows:

Stage 1: select the leader from the n people, $\binom{n}{1} = n$ ways to complete this. Then

Stage 2: select a subset from the remaining $n - 1$ members to round out the club, 2^{n-1} ways.

The basic counting rule says there are

$$n2^{n-1} \tag{1}$$

clubs.

Here's *another* solution where we first fix the size k of the club, count the number of clubs of size k, and then sum from $k = 1$ to $k = n$ to get all the clubs. To this end, fix k between 1 and n (inclusive). Here's a process that leads to a club of size k:

Stage 1: select a subset of size k from the n possible, $\binom{n}{k}$ ways. Then

Stage 2: identify one the people in the subset in stage 1 as the leader, $\binom{k}{1} = k$ ways.

The basic counting rule says there are $k\binom{n}{k}$ clubs of size k. Therefore, there are

$$\sum_{k=1}^{n} k\binom{n}{k} \tag{2}$$

clubs.

The answers (1) and (2) look different, but if we did the counting right, they should be the same.

Exercise for the student[†]**.**
Show that

$$\sum_{k=1}^{n} k\binom{n}{k} = n2^{n-1}.$$

[†] This is a nice (and, if you are doing it right) short calculation. Seriously, do this ***now***! Hint: you'll need to recognize to use the corollary to the binomial theorem.

We now collect some important mathematical results involving binomial coefficients.

• **FACT 1**: For any integers $n \geq 1$ and $0 \leq k \leq n$,

$$\binom{n}{k} = \binom{n}{n-k}.$$

Here's a very simple algebraic proof:

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

Here's a very simple combinatorial proof:
From a set of n people, we plan to select k winners and the remaining $n-k$ will be losers. $\binom{n}{k}$ represents the number of subsets of k winners. Since every subset of k winners we select corresponds to a unique subset of $n-k$ losers, the subsets of k winners is in one-to-one correspondence with the subsets of $n-k$ losers. Therefore, $\binom{n}{k} = \binom{n}{n-k}$.

FACT 2: (**Pascal's identity**)
For any integers $n \geq 1$ and $0 < k < n$ (i.e., $1 \leq k \leq n-1$),

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

Here's an algebraic proof:
Let n and k be integers as described in the statement.

$$\begin{aligned}
\binom{n-1}{k} + \binom{n-1}{k-1} &= \frac{(n-1)!}{k!(n-1-k)!} + \frac{(n-1)!}{(k-1)!(n-k)!} \\
&= \frac{(n-1)!(n-k)}{k!(n-1-k)!(n-k)} + \frac{(n-1)!\,k}{(k-1)!(n-k)!\,k} \\
&= \frac{(n-1)!\,n-(n-1)!\,k}{k!(n-k)!} + \frac{(n-1)!\,k}{k!(n-k)!} \\
&= \frac{n!}{k!(n-k)!} = \binom{n}{k}.
\end{aligned}$$

EXERCISES.

1. In how many ways can a coach create a tee-ball team of 9 players from a collection of 15 players?

2. Harry buys 5 cookies from Insomnia Cookie. In how many ways can he create the box so that all the flavors are distinct if there are 40 different flavors?

3. A barber is creating a fresh fade. He has to choose 4 different scissors from a collection of 16 scissors.
In how many ways can the barber select his scissors?

4. George is selecting 6 different fencing uniforms to take with him on a trip. He has 24 uniforms total. How many different selections of uniforms can George take with him?

5. Vincent has all 9 trophies from High Tide in his room. He wants to move 4 of them to Simon's room. In how many ways can Vincent select 4 trophies to give to Simon?

6. In Spikeball, there are 8 valid starting positions around the circular net that one can stand in. Four (4) people are playing Spikeball. In how many ways can they select 4 spots out of these 8 to be occupied at the start if we only care about the positions of the people relative to each other?

7. Miguel has 20 different earrings that he wears regularly. In how many ways can he select two to wear on a given day?

8. Brandon is running a Hot Dog Joint that has 32 distinct orders. In how many ways can a customer purchase 8 distinct items from the menu?

9. How many binary sequences (sequences only consisting of 0 and 1) of length 14 have exactly 6 ones?

10. Kevin works at the Applied Physics Laboratory. There are 21 projects being worked on, and he must choose 4 to supervise. In how many ways can Kevin do this?

11. For each integer $n \geq 1$, simplify $\sum_{k=0}^{n}(-1)^k\binom{n}{k}$. What does this result say combinatorially about a set of size n?

12. Compute this sum: $\displaystyle\sum_{k=0}^{n}\frac{1}{k!(n-k)!}$. You should get a function of n alone.

13. If $(2x-1)^{10}$ is written out as a polynomial in x of degree 10, determine the coefficient of x^8 in this expansion. Then do x^5.

14. A dissertation defense committee at Johns Hopkins is a group of 5 people: one is the student's dissertation advisor, 3 are eligible members of the faculty in the advisor's department, and an eligible faculty from outside the advisor's department. Rhee Lee Smart is trying to form her dissertation committee. Her dissertation advisor is Justin Case from the Department of Civil Disobedience (DOCD). There are 600 other eilgible university faculty that can serve but only 8 of these belong to DOCD. How many dissertation committees can Rhee form?

15. From a pack a 20 m&m's there are 5 red, 4 blue, 3 green, 6 yellow, and 2 orange. Assume the candies are well-mixed. Only simplify if it's something nice.
These are separate questions unless noted otherwise.
(a) We grab a handful of 4 m&m's from this pack. What's the probability that you grab exactly 2 red m&m's?
(b) The plan is to line up all 20 m&m's. What's the chance that no two red m&m's are adjacent?
(c) (continued from part (b)) What's the chance the exactly two red m&m's are adjacent?

16. Roll a fair 6-sided die 5 times. Compute the probability you throw a two-pair.

17. I deal you 8 cards from a (well-shuffled) standard deck of 52. What's the probability that you get exactly 2 of each suit?

Somewhat challenging (?) problems.

18. A is a finite set with $n > 1$ members. We plan to select a subset of A at random but *not* uniformly at random. In fact, we are told that if $|\omega| = k$, then $P(\{\omega\}) = \dfrac{k}{n2^{n-1}}$. Notice $P(\{\varnothing\}) = 0$.
(a) S_k is the event that we select a subset of size k from A, $0 \leq k \leq n$. Compute $P(S_k)$.
(b) Show that $\displaystyle\sum_{\omega\in\Omega} P(\omega) = 1$.

Remark. It might help to first do this problem in the special case where $A = \{1, 2, 3\}$ and $\Omega = \{\varnothing, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.

19. We turn over the cards of a well-shuffled standard deck of 52 cards one by one. What's the probability that all the Kings are turned over before the first Ace? What's the probability that exactly two Kings are turned over before the first Ace?

Counting with multinomial coefficients.

Consider the following problem:

We have n distinct objects.
We have r distinct "boxes" labeled 1 thru r.
We want to put the n objects into the boxes with the following restrictions:

1. Each object is put into only one box.
2. The order in which objects are put into boxes is irrelevant.
3. We have prescribed nonnegative integers $n_1, n_2, \ldots, n_r$ such that

$$n_1 + n_2 + \cdots + n_r = n,$$

that tell us how many objects to put in each "box": box i gets n_i of the objects.
We ask:

How many ways can we assign the objects to these boxes?

Let's start with, say, box 1. We need to put n_1 objects into box 1 and order is irrelevant. So, select a subset of n_1 objects for box 1. Now, there are $n - n_1$ objects remaining. Now, say, go to box 2, which is to receive n_2 objects. Select a subset of n_2 objects from the $n - n_1$ remaining for box 2, and so on.

This scheme creates an r-stage procedure that fits the paradigm of the basic counting rule:

Stage 1: select n_1 for box 1.
Stage 2: select n_2 from those remaining (after the first stage) for box 2.
Stage 3: select n_3 from those remaining (after the first 2 stages) for box 3.
$\vdots$
Stage $r-1$: select n_{r-1} from those remaining (after first $r-2$ stages) for box $r-1$.
Stage r: there are now n_r remaining, just put these in r.

Here's the resulting count:

$$\underbrace{\binom{n}{n_1}}_{\frac{n!}{n_1!(n-n_1)!}} \cdot \underbrace{\binom{n-n_1}{n_2}}_{\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}} \cdot \underbrace{\binom{n-n_1-n_2}{n_3}}_{\frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!}} \cdots \underbrace{\binom{n-n_1-n_2-\cdots-n_{r-2}-n_{r-1}}{n_r}}_{\frac{(n-n_1-\cdots-n_{r-1})!}{n_r!(n-n_1-\cdots-n_{r-1}-n_r)!} \,=\, 1}.$$

You may notice that each binomial coefficient in this product from the second one onward has the property that the factorial in the numerator appears as the denominator in the previous binomial coefficient. So, there's a significant amount of cancellation (a telescoping product, in fact), and I leave this...

Exercise for the student.
Show that the product of binomial coefficients above reduces to* $\dfrac{n!}{n_1!n_2!\cdots n_r!}$.

* This is the number of anagrams of an n-letter word with n_i letters of type i, i from 1 to r.

Notation.
We may use the symbol

$$\binom{n}{n_1, n_2, \ldots, n_r}$$

to represent $\dfrac{n!}{n_1!n_2!\cdots n_r!}$, and refer to it as a ***multinomial coefficient***.

Remark. (the multinomial coefficient is counting *ordered* r-tuples)
It's worth pointing out the sample points that the multinomial coefficient is counting are *ordered* r-tuples of subsets with specified sizes for each entry, where the r subsets are pairwise disjoint and union to the entire set of n objects. In this context we also assume these subsets are *nonempty*, because if a box is to receive no objects we can just remove the box from the experiment.

I'll also mention that because we can distinguish between the boxes, if 2 subsets of the same size in a r-tuple are swapped, then this leads to a different assignment.

Example.
We have 6 balls numbered 1 thru 6 that are to be put into 3 boxes, where box 1 gets 3, box 2 gets 2, and box 3 gets 1 ball. How many ways can this be done?

SOLUTION:

There are $\binom{6}{3, 2, 1} = \dfrac{6!}{3!2!1!} = 60$ ways.

In this example, the boxes are clearly distinguishable from each other because each box is getting a different number of balls: box 1 is the one getting 3 balls, box 2 is the one getting 2 balls, and box 3 is the one getting 1 ball.

Example.
A teacher has 6 children, a, b, c, d, e, f, in a kindergarten class. She plans to assign each child a task. There are 3 tasks:

task 1: move desks in a circle
task 2: throw out the trash
task 3: erase the blackboard.

The teacher plans to assign 3 to task 1, 2 children to task 2, and 1 child to task 3.
How many ways can the teacher assign the children to a task (each child gets exactly one task)?

SOLUTION: This is the same as the previous example, $\binom{6}{3, 2, 1}$. (*Do you see why?*)

Again, in this problem we can clearly distinguish between the tasks(boxes) since each task requires different number of children.

Remark.
The multinomial coefficient $\binom{n}{n_1, n_2, \ldots, n_r}$ counts the number of ways we can assign subsets of objects to distinct boxes – putting subsets of objects to boxes - where each box gets a prescribed number of objects: n_i to box i. But, from the formula at the top of page 42, the multinomial coefficient is also the number of anagrams of an n-letter word, where there are n_i *indistinguishable* letters of type i – we are mapping the n positions(objects) to r letters(boxes) . Here's the connection between these two approaches...

In the last example, let's call the tasks D (for "move the desks"), T (for "throw out the trash"), and B (for "erase the blackboard"), and fix a line-up of the 6 children left to right, say we choose (a, b, c, d, e, f) as our line-up. Create a 6-letter word with 3 D's, 2 T's, and 1 B. For example, the word $DDDTTB$, thought of as

$$\frac{D}{a}\frac{D}{b}\frac{D}{c}\frac{T}{d}\frac{T}{e}\frac{B}{f},$$

is telling us that a, b and c are being assigned to job D, d and e are being assigned to job T, and f is being assigned to job B.

Below we demonstrate the correspondence: the ordered 3-tuple is sending subsets of children to tasks and the anagram is mapping of children to tasks. The two approaches are just different ways to think about the same thing.

$$\begin{aligned}\Big(\{a,b,c\},\{d,e\},\{f\}\Big) &=: \{a,b,c\} \mapsto D, \{d,e\} \mapsto T, \{f\} \mapsto B \\ DDDTTB &=: a \mapsto D, b \mapsto D, c \mapsto D, d \mapsto T, e \mapsto T, f \mapsto B\end{aligned}$$

$$\begin{aligned}\Big(\{b,d,f\},\{c,e\},\{a\}\Big) &=: \{b,d,f\} \mapsto D, \{c,e\} \mapsto T, \{a\} \mapsto B \\ BDTDTD &=: a \mapsto B, b \mapsto D, c \mapsto T, d \mapsto D, e \mapsto T, f \mapsto D\end{aligned}$$

$$\begin{aligned}\Big(\{b,c,e\},\{a,f\},\{d\}\Big) &=: \{b,c,e\} \mapsto D, \{a,f\} \mapsto T, \{d\} \mapsto B \\ TDDBDT &=: a \mapsto T, b \mapsto D, c \mapsto D, d \mapsto B, e \mapsto D, f \mapsto T \\ &\vdots\end{aligned}$$

Example.
In the card game *Bridge* there are 4 people called North, South, East, West and a standard deck of 52 cards. A *Bridge deal* is when we deal 13 cards to each of these 4 people – so that there are four (4) 13-card hands being dealt. For this problem assume the order in which the 13 cards are dealt to each person is irrelevant. Moreover, if, for instance, everyone took their hand and rotated to the person on the right then that would be considered a *different* Bridge deal. How many Bridge deals are possible?

SOLUTION:
Since we distinguish the players hands, the way we view this is like this: the dealer deals a 13-card hand to, say, North; then, from the remaining $52 - 13 = 39$ cards, she deals a 13-card hand to, say, South; and, so on. A 4-stage procedure! We have

$$\binom{52}{13} \cdot \binom{39}{13} \cdot \binom{26}{13} \cdot \binom{13}{13}.$$

Similarly, 52 distinct objects, 4 boxes and 13 objects to each box. Answer: $\binom{52}{13, 13, 13, 13}$.

ADVICE: For a Bridge deal it is sometimes helpful to think of the sample points as the ordered 4-tuple of 13-card hands:

$$\Big(\underbrace{\{\text{13-card hand}\}}_{\text{North's hand}}, \underbrace{\{\text{13-card hand}\}}_{\text{South's hand}}, \underbrace{\{\text{13-card hand}\}}_{\text{East's hand}}, \underbrace{\{\text{13-card hand}\}}_{\text{West's hand}} \Big).$$

Example.
How many Bridge deals have North receiving all the Aces?

SOLUTION:
Every hand North receives will now look like $\{\text{A}\clubsuit, \text{A}\diamondsuit, \text{A}\heartsuit, \text{A}\spadesuit\} \cup \{\text{9-card hand}\}$, and there are no restrictions on the 13-card hands of the other three players. If we imagine listing out all of North's hands, you will see that the only differences are the subset of size 9 that are chosen from the 48 non-Ace cards available. So counting the number of ways North can receive all the Aces is the same as taking the 4 Aces out of the original deck and dealing North only 9 cards from the 48, and 13 cards for each of the other players. The answer is

$$\binom{4}{4}\binom{48}{9} \cdot \binom{39}{13} \cdot \binom{26}{13} \cdot \binom{13}{13} = \binom{48}{9, 13, 13, 13}.$$

Alternatively, counting the number of hands where North receives all the Aces can be thought of as a 2-stage process: select the 4 Aces, $\binom{4}{4} = 1$ way to do this; then, select the other 9 cards from 48, $\binom{48}{9}$. The rest of the counting is the same as above.

All the examples thus far in this section have fit the paradigm of the multinomial coefficient: n distinct objects going into r distinct boxes with n_i objects for box i ($i = 1, 2, \ldots, r$. Here's an example where we can't distinguish between the boxes...

Example and discussion.
A teacher has 6 children, a, b, c, d, e, f, in a kindergarten class. She plans to create 3 teams of size 2. How many ways can the teacher create the teams?

SOLUTION:
There is now a subtlety that we should discuss. Had we said the 3 teams were the A-team, the B-team, and the C-team, then these teams are clearly labeled. We'd have to form the assignment by saying which 2 children are the A-team, which 2 are the B-team, and which 2 are the C-team, and we can distinguish between assignments because the teams are clearly labeled.

But, in our current problem, the teams are *not* labeled; moreover, they are all the same size, so we cannot distinguish between the teams by their size either. The teams are now *indistinguishable*! Thus, for instance,

$$\Big(\{a,d\},\{b,f\},\{c,e\}\Big),\quad \Big(\{a,d\},\{c,e\},\{b,f\}\Big),\quad \Big(\{b,f\},\{a,d\},\{c,e\}\Big),$$

$$\Big(\{b,f\},\{c,e\},\{a,d\}\Big),\quad \Big(\{c,e\},\{a,d\},\{b,f\}\Big),\quad \Big(\{c,e\},\{b,f\},\{a,d\}\Big)$$

all represent the same 3 teams. i.e., they are all the same assignment. So, in the current problem we need to ignore the order in which we place the subsets in the assignment. Counting the number of assignments as if these were really an A-team, a B-team, and a C-team would be *over-counting* by a factor of 3! since, for any ordered 3-tuple, any rearrangement of the subsets will lead to the *same* assignment – and there are 3! ways to order the 3 distinct subsets within the 3 positions. Consequently, there are

$$\frac{\binom{6}{2,2,2}}{3!} = 15$$

ways for the teacher to create 3 teams of size 2 each from the 6 children.

Remark.
If the "boxes" all get differing numbers of objects in each, then we can distinguish between the boxes by the number of objects in them. When some (or all) of the subsets being assigned to the boxes are the same size, however, we should be a bit careful and think about the situation. Usually, whether or not boxes are distinguishable is clear from context. Hopefully the last example demonstrated this.

Example.
We divide up a standard deck of 52 cards into 4 piles of 13 cards. How many divisions are possible?

SOLUTION:
This is almost the same question as the Bridge deals problem from earlier, except now the piles are indistinguishable. Visualize yourself doing this, then ask yourself: If, after dividing the deck up twice and comparing the resulting divisions (sample points), I find that the piles each have exactly the same cards in each, do I *really* want to say these are different piles? The answer should be *probably not*! There is nothing in this question that suggests we need to distinguish the piles from each other. The answer would then be

$$\frac{\binom{52}{13,13,13,13}}{4!}$$

divisions of the deck into 4 equal-sized piles.

Example and discussion.
We need to assign 10 people to 4 distinct jobs – each person gets one job. Jobs 1, 2, and 3 require 2 people each while job 4 requires 4 people. How many assignments of jobs to people are there?

SOLUTION: There are $\binom{10}{2,2,2,4} = \frac{10!}{2!2!2!4!} = 18900$ assignments possible.

Let's think for a minute about the objects that we are counting. If we call the 10 people $a, b, c, d, e, f, g, h, i, j$, then we can think of an assignment as an ordered 4-tuple of subsets like this

$$\Big(\underbrace{\{d,g\}}_{\text{job 1}}, \underbrace{\{a,h\}}_{\text{job 2}}, \underbrace{\{b,i\}}_{\text{job 3}}, \underbrace{\{c,e,f,j\}}_{\text{job 4}}\Big).$$

The subsets of people assigned to each job are disjoint – no one is assigned more than one job – and, their union is all 10 people – everyone is assigned a job.

Suppose the first 3 jobs (having 2 people in each) are really the same job, say, cleaning windows. What changes with the answer above?

Now, the same subsets of people assigned to jobs $1, 2$, and 3 in a different order is the same assignment. So, we get

$$\frac{\binom{10}{2,2,2,4}}{3!}$$

instead.

EXERCISES.

1. How many ways can 30 distinct objects be divided into 3 distinct subsets of respective sizes 16, 8 and 6?

2. 11 girls are to be assigned soccer positions (one girl to each position):
1 forward, 5 midfielders, 4 defenders, and 1 goalie.
How many different assignments are possible?
How many assignments having Angela as a defender are possible?

3. Gary is creating a workout. The order of the exercises he performs is irrelevant. Out of the 28 machines, in how many ways can he select 4 machines to do each day of the week with no repeats?

4. Jan is packing up his 18 math textbooks into 3 boxes. In how many ways can he do this if the 3 boxes hold $4, 6$, and 8 books, respectively?

5. In how many ways can 15 students be split into 3 groups to work on a project? Assume there is at least one person in each group.

6. How many distinct arrangements of FREDTORCASOPROB are there?

7. A stack of 18 dinner plates is created with 4 red, 4 blue, 5 green, and 5 yellow plates. How many distinct stacks are possible?

8. Harry orders 1 of each of the 40 flavors Insomnia cookie has and then puts them into 8 boxes of 5 cookie capacities. In how many ways can Harry arrange the cookies?

9. Dan buys 15 Pokemon Card packs, of which 6 are Shining Fates, 4 are Roaring Skies, 3 are Sword and Shield, and 2 are Celebrations. He wants to open all 15 packs. In how many distinguishable orders can he do this?

10. A 6-sided die is rolled $6n$ times. How many sequences of rolls have exactly n of each of the 6 values?

11. Gabe places an order from Billabong. He purchases 12 shirts: 3 Wrangler Series, 3 Simpsons Collaboratory, 3 50th Anniversary, and 3 Basic. He wants to wear these 12 shirts on 12 consecutive days. In how many distinct ways can Gabe wear the shirts in a 12-day period, assuming he is cleanly and, therefore, wears each shirt exactly once over this period?

Stars-and-bars

The experiment here is a slight modification to the sampling with replacement scheme, and we'll get sample points that we haven't quite seen yet.

The stars-and-bars experiment.
We have a set of r distinct labels. We sample n of these labels with replacement, but *ignore the order in which they were drawn.*

Basic example.
If the $r = 3$ labels were $\{1, 2, 3\}$ and we sample $n = 2$ with replacement – ignoring the order in which we selected them, then I'll write the possible sample points like this:

$$\{1, 1\}, \{2, 2\}, \{3, 3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}.$$

Here, $\{1, 1\}$ means we drew label 1 twice, $\{1, 3\}$ means we drew label 1 and label 3 (each once), etc. There are 6 sample points total. Please convince yourself that we're not missing any. I stress that, since we are ignoring order, the sample point $\{1, 2\}$, for instance, is the same as $\{2, 1\}$ and it only needs to be represented once.

Remark.
WARNING: Please do *not* view the sample points in this example as sets, because, if we did, then $\{1, 1\} = \{1\}$ and we'd lose the interpretation that the label 1 appears $n_1 = 2$ times. In this context, we view the sample points above as *multisets*. As the example illustrates, we can think of a multiset as an *unordered n-tuple of labels 1 thru r, where repetition is allowed.*

Multisets.
A ***multiset*** is a set that allows labels to appear more than once. The ***multiplicity*** (or ***index***) of a label in a multiset is the number of times the label appears in it. We may use the notation n_i to represent the multiplicity of label i ($i = 1, 2, \ldots, r$). We allow $n_i = 0$, which will mean the label i does not appear in the multiset (i wasn't selected). The ***size*** of a multiset is the total number of labels in it including multiplicity; in the stars-and-bars experiment above we are drawing n labels, so the resulting size of the multiset will be n. In this way, two multisets are ***equal*** if each label in the multisets have the same multiplicity, which will imply they are the same size. For instance, among the following multisets:

$$\{1, 1, 1, 2, 2, 4\}, \quad \{1, 2, 4\}, \quad \{1, 1, 1, 2, 2, 3, 4\}, \text{ and } \{4, 2, 2, 1, 1, 1\}$$

only the first and the last ones are equal since they are the only ones that represent the same unordered 6-tuple, and also because $n_1 = 3$, $n_2 = 2$, $n_3 = 0$, and $n_4 = 1$ for each of them. Notice this also implies that they will have size $n_1 + n_2 + n_3 + n_4 = 6$.

Just to clarify: the stars-and-bars experiment above can be viewed as producing sample points that are unordered n-tuples of r distinct objects (repetition allowed).

There is an alternate way of viewing the stars-and-bars experiment that will lead to an equivalent concept to a multiset. This equivalent way of viewing the sample space will provide us with a formula to count the total number of multisets for a given n and r.

Alternate way to represent sample points in the stars-and-bars-experiment.
Let n and r be positive integers. Without loss of generality, let's assume the set of r labels we are selecting from is $\{1, 2, \ldots, r\}$. When we perform the stars-and-bars experiment, we get a multiset of size n of (possibly repeated) labels. Instead of recording the multiset, we could, equivalently, record the ordered r-tuple of the multiplicities of each label, i.e., record $(n_1, n_2, \ldots, n_r)$. For a fixed n and r, there's a one-to-one correspondence between multisets and these ordered r-tuples $(n_1, n_2, \ldots, n_r)$ of nonnegative integers with $n_1 + n_2 + \cdots + n_r = n$. The ordered r-tuple is an ordered ***integer partition of*** n.

Illustration and important FACT.
Take $n = 3$ and $r = 4$. These numbers are relatively small so I can attempt to list all the resulting multisets, but I think you can see how listing these will become tedious had n and r been larger. To the right of each multiset I put the corresponding integer partition. Ignore, for the moment, the stars-and-bars representations (I will discuss on page 51 but in the meantime if you see a pattern that's great!).

multiset		integer partition		stars-and-bars
$\{1,1,1\}$	$\longleftrightarrow$	$(3,0,0,0)$	$\longleftrightarrow$	$\star\star\star\vert\vert\vert$
$\{2,2,2\}$	$\longleftrightarrow$	$(0,3,0,0)$	$\longleftrightarrow$	$\vert\star\star\star\vert\vert$
$\{3,3,3\}$	$\longleftrightarrow$	$(0,0,3,0)$	$\longleftrightarrow$	$\vert\vert\star\star\star\vert$
$\{4,4,4\}$	$\longleftrightarrow$	$(0,0,0,3)$	$\longleftrightarrow$	$\vert\vert\vert\star\star\star$
$\{1,1,2\}$	$\longleftrightarrow$	$(2,1,0,0)$	$\longleftrightarrow$	$\star\star\vert\star\vert\vert$
$\{1,1,3\}$	$\longleftrightarrow$	$(2,0,1,0)$	$\longleftrightarrow$	$\star\star\vert\vert\star\vert$
$\{1,1,4\}$	$\longleftrightarrow$	$(2,0,0,1)$	$\longleftrightarrow$	$\star\star\vert\vert\vert\star$
$\{2,2,3\}$	$\longleftrightarrow$	$(0,2,1,0)$	$\longleftrightarrow$	$\vert\star\star\vert\star\vert$
$\{2,2,4\}$	$\longleftrightarrow$	$(0,2,0,1)$	$\longleftrightarrow$	$\vert\star\star\vert\vert\star$
$\{3,3,4\}$	$\longleftrightarrow$	$(0,0,2,1)$	$\longleftrightarrow$	$\vert\vert\star\star\vert\star$
$\{1,2,2\}$	$\longleftrightarrow$	$(1,2,0,0)$	$\longleftrightarrow$	$\star\vert\star\star\vert\vert$
$\{1,3,3\}$	$\longleftrightarrow$	$(1,0,2,0)$	$\longleftrightarrow$	$\star\vert\vert\star\star\vert$
$\{1,4,4\}$	$\longleftrightarrow$	$(1,0,0,2)$	$\longleftrightarrow$	$\star\vert\vert\vert\star\star$
$\{2,3,3\}$	$\longleftrightarrow$	$(0,1,2,0)$	$\longleftrightarrow$	$\vert\star\vert\star\star\vert$
$\{2,4,4\}$	$\longleftrightarrow$	$(0,1,0,2)$	$\longleftrightarrow$	$\vert\star\vert\vert\star\star$
$\{3,4,4\}$	$\longleftrightarrow$	$(0,0,1,2)$	$\longleftrightarrow$	$\vert\vert\star\vert\star\star$
$\{1,2,3\}$	$\longleftrightarrow$	$(1,1,1,0)$	$\longleftrightarrow$	$\star\vert\star\vert\star\vert$
$\{1,2,4\}$	$\longleftrightarrow$	$(1,1,0,1)$	$\longleftrightarrow$	$\star\vert\star\vert\vert\star$
$\{1,3,4\}$	$\longleftrightarrow$	$(1,0,1,1)$	$\longleftrightarrow$	$\star\vert\vert\star\vert\star$
$\{2,3,4\}$	$\longleftrightarrow$	$(0,1,1,1)$	$\longleftrightarrow$	$\vert\star\vert\star\vert\star$

Hopefully, this illustration convinces you of the following

FACT: the total number of multisets of size n from a set of r labels *is the same as* the total number of nonnegative integer r-tuple $(n_1, n_2, \ldots, n_r)$ solutions to the equation

$$n_1 + n_2 + \cdots + n_r = n.$$

Counting one thing is the same as counting the other.

The fact on the previous page is important because the ordered r-tuple interpretation motivates a very clever idea/insight into counting the total number of multisets (and, thus, nonnegative integer r-tuple solutions to $n_1 + n_2 + \cdots + n_r = n$).

Stars-and-bars counting.
Fix positive integers n and r. *How do we count the number of possible multisets?*
The idea behind this counting dates back to the time of Max Planck but popularized by William Feller in his classic probability book, *Probability Theory and its Applications, Volume 1.* From the fact, we can count the number of multisets by instead counting the number of ordered r-tuples of the form

$$(n_1, n_2, \ldots, n_r),$$

where each entry is a nonnegative integer and all entries sum to n. The multisets we are counting have n labels. Abstractly, represent each label by a **star** $\star$, so that we have n stars. In the representation of the r-tuple there are $r-1$ commas that separate (partition) the nonnegative integers $n_1, n_2, \ldots, n_r$ into r pieces. Let's represent a comma in the r-tuple by a ***bar*** $|$, so that we have $r-1$ bars. Interpret the ordered r-tuple above as having

n_1 stars to the left of the bar 1,
n_2 stars between bar 1 and bar 2,
n_3 stars between bar 2 and bar 3,
$\vdots$
n_r stars to the right of bar $r-1$.

We, therefore, have represented the ordered r-tuple as a "word" involving stars and bars:

$$\underbrace{\star\star\cdots\star}_{n_1 \text{ stars}}|\underbrace{\star\star\cdots\star}_{n_2 \text{ stars}}|\underbrace{\star\star\cdots\star}_{n_3 \text{ stars}}|\cdots\cdots|\underbrace{\star\star\cdots\star}_{n_r \text{ stars}}$$

Important observation:
Each anagram of this $(n+r-1)$-letter word corresponds to a unique ordered r-tuple (i.e., ordered integer partition of n) and, therefore, to a unique multiset. Thus, the number of anagrams of this word *is* the total number of multisets of size n from r distinct objects. The word has n $\star$'s and $r-1$ $|$'s, so the answer is given in the following result.

Counting multisets.
The expression

$$\frac{(n+r-1)!}{n!(r-1)!} = \binom{n+r-1}{n} = \binom{n+r-1}{r-1}$$

counts

• the number of unordered n-tuples of r distinct labels repetition allowed,

which is the same as

• the number of ordered r-tuples $(n_1, n_2, \ldots, n_r)$ of nonnegative integer solutions to

$$n_1 + n_2 + \cdots + n_r = n.$$

Remark.
In the illustration on page 49 I show the correspondence between the stars-and-bars to the multisets and to the ordered r-tuples. The $r-1=4-1=3$ bars left-to-right create 4 ordered buckets left-to-right. The bucket to the left of the first bar, the bucket between the first and second bar, the bucket between the second and third bar, and finally, the bucket to the right of the third bar. So, for instance, in the first one: $\star\star\star|||$ says

$$\underbrace{\star\star\star}_{3}|\underbrace{\ }_{0}|\underbrace{\ }_{0}|\underbrace{\ }_{0},$$

the representation $|\star||\star\star$ says

$$\underbrace{\ }_{0}|\underbrace{\star}_{1}|\underbrace{\ }_{0}|\underbrace{\star\star}_{2},$$

and, so on.

We will now do several examples.

Example.
How many ways can we distribute 8 *identical* balls into 5 *distinct* boxes?

SOLUTION:
Viewing this as an integer partition problem: we can observe how many identical balls went into each box. I.e., we are counting the number of nonnegative integer solutions $(n_1, n_2, n_3, n_4, n_5)$ to $n_1+n_2+n_3+n_4+n_5=8$. Of course, n_i is the number of identical balls in box i. The answer is

$$\frac{(8+5-1)!}{8!(5-1)!}=\frac{12!}{8!4!}=495.$$

On the other hand, we could have viewed the counting above as a multiset problem as follows: assign one of the labels 1 thru 5 to each of the 8 balls allowing labels to be repeated. The label on a ball will tell us which box that ball is in. But, we cannot distinguish balls – they are identical. So, if the 8 balls received the labels like this:

$$1,4,4,4,5,3,2,2$$

and also like this:

$$4,4,4,2,2,1,3,5$$

then we cannot tell the difference because, again, balls are indistinguishable: we can't say, for instance, that in the first assignment ball 1 got label 1, but in the second assignment, ball 1 got label 4. How would you know which one ball 1 is?? They aren't distinguishable. So, we are really counting unordered 8-tuples of the numbers 1 thru 5 repetition allowed, i.e., from this point of view we are counting multisets.

Example and discussion.
Imagine we are forming bags of 10 marbles. The marbles come in 5 different colors, but are otherwise identical. We can fill our bag with as many or as few of each color we desire[2]. How many different bags of marbles can be formed?

SOLUTION:
How should we start? If you're stuck, write down what a "few" bags might look like. Say, one bag is all 10 color 1's. Another bag is 9 color 1's and 1 color 2. Yet another is 2 of color i, $i = 1, 2, 3, 4, 5$. And, there are *many* more! Each bag will always have 10 marbles, but the numbers of each color will vary from bag to bag. How should we represent a sample point? What model can we choose for a *bag* of such marbles? After listing a few bags we get an idea for a model. Each bag can be identified by how many of each color went into it. So, we can represent such a bag as an ordered 5-tuple $(n_1, n_2, n_3, n_4, n_5)$, where n_i is the number of color i marbles in the bag. We need $n_1+n_2+n_3+n_4+n_5 = 10$. The answer is

$$\frac{(10+5-1)!}{10!(5-1)!} = \frac{14!}{10!4!} = 1001 \text{ bags.}$$

Example and discussion.
Throw 5 *identical* 6-sided dice simultaneously. How many outcomes are possible?

SOLUTION:
Of course, we know the answer if these dice were distinguishable: 6^5 because, in this case, we would be counting *ordered* 5-tuples where each entry is any of the ranks 1 thru 6 repetition allowed; maybe the distinguishable dice are different colors that we can plainly see, or we are thinking that the throws of the identical dice are one-at-a-time and we observe the order of the ranks... all to justify the equally likely assumption. But, the situation now is *not* this. So, what do we observe? What's a model for the sample points?

One point of view is this: since the dice are indistinguishable, then all we observe is the *unordered* 5-tuples of the numbers 1 thru 6 repetition allowed, i.e., we observe multisets with $n = 5$. Moreover, the entries of these multisets are ranks belonging to the set $\{1, 2, 3, 4, 5, 6\}$, this implies $r = 6$. So, the number of multisets is

$$\frac{(5+6-1)!}{5!(6-1)!} = \frac{10!}{5!5!} = \binom{10}{5} = 252,$$

far fewer than $6^5 = 7776$.

Alternatively, if we view the experiment as throwing indistinguishable dice all at once, then all we can observe is how many (identical) dice are showing each rank. I.e., we observe n_i dice showing rank i, $i = 1, 2, 3, 4, 5, 6$. And, $n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 5$. So, the number of possible ordered 6-tuple solutions, $(n_1, n_2, n_3, n_4, n_5, n_6)$, is, again,

$$\frac{(5+(6-1))!}{5!(6-1)!}.$$

[2]We assume that we have an unlimited supply of marbles of each color.

Thought exercise.
In the last example, should we view these 252 sample points as equally likely? To get a feel for an answer, re-do this example with only 2 indistinguishable dice instead of 5. In this case, interpret what having these outcomes being equally likely means compared to experiment 2.2 on page 7.

Remark.
The last few examples show the power of star and bars counting when we can model the sample points as multisets of size n, i.e., as unordered n-tuples of r distinct labels repetition allowed, or as ordered r-tuples of nonnegative integers that sum to a positive integer n. Indeed, the sample space of all such sample points has cardinality

$$\frac{(n+r-1)!}{n!(r-1)!}.$$

Now for some slight generalizations...

Example.
Throw 5 identical 6-sided dice. How many outcomes will show *exactly* one 6?

SOLUTION:
Now, we are being told that we want the number of nonnegative integer solutions to

$$n_1+n_2+n_3+n_4+n_5+n_6=5 \qquad \textbf{\textit{and}} \qquad n_6=1.$$

This means we are looking for the number of nonnegative integer solutions to

$$n_1+n_2+n_3+n_4+n_5=4.$$

Here, $n=4$ and $r=5$, so there are

$$\frac{(4+(5-1)!}{4!(5-1)!}=\frac{8!}{4!4!}=70$$

throws of 5 identical dice that show exactly one 6.

Here's the multiset way of looking at the problem: we are trying to count all the multisets that have exactly one 6. The structure of these multisets is rather simple; they look like

$$\{6\}\cup\{\text{a multiset of size 4 of labels 1,2,3,4,5}\}.$$

These multisets are, clearly, in one-to-one correspondence with multisets of size 4 drawn from 5 labels, and we'll get the same answer as above. The intuition this point of view gives is this: put exactly one rank 6 into each multiset, and then count the number of multisets of one less size from the remaining ranks.

Let's generalize more...

Example.
Throw 5 identical 6-sided dice. How many outcomes will show a 6?

SOLUTION:
An outcome showing a 6 means that a 6 occurs on the throw of 5 dice, which implies that *at least* one 6 appears on the throw. What does this constraint do to the counting we did in the last example? Now, it is clear that we want to count the number of nonnegative integer solutions to

$$n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 5 \qquad \textbf{\textit{and}} \qquad n_6 \geq 1.$$

Now, $n_6 \geq 1$ is the same as $\widetilde{n}_6 := n_6 - 1 \geq 0$. Then the above becomes

$$n_1 + n_2 + n_3 + n_4 + n_5 + n_6 - 1 = 5 - 1 \qquad \textit{and} \qquad \widetilde{n}_6 := n_6 - 1 \geq 0,$$

or, equivalently, we are looking for the number of nonnegative integer solutions to

$$n_1 + n_2 + n_3 + n_4 + n_5 + \widetilde{n}_6 = 4.$$

But this now fits the paradigm of the stars-and-bars counting with $n = 4$, $r = 6$:

$$\frac{(4 + (6-1))!}{4!(6-1)!} = \frac{9!}{4!5!} = 126$$

throws where a 6 occurs.

Thought exercise.
In this last example, what would be the reasoning using the multiset point of view?

EXERCISES.

1. In how many ways can you give 9 children 14 chocolate chip cookies?

2. Repeat Question 1 so that Rick, one of the kids, receives exactly one cookie.

3. Repeat Question 1 so that Rick receives at least two cookies.

4. Repeat Question 1 so that no child goes hungry (i.e., each receives at least one cookie).

5. Blaze Pizza offers 12 choices of toppings for their pizzas. You can get none of a topping, one of a topping, double, triple, etc, but, for instance, double sausage counts will count as two toppings. How many 8 topping pizzas are possible?

6. A *multiset* is a set that is allowed to have repeats. For example, $\{1, 2, 5, 5, 7, 7\}$ is a multiset of size 6. How many multisets of size 6 are there that consist of the digits 0 thru 9?

7. A PE teacher puts 20 dodgeballs away into 6 bins. In how many ways can the PE teacher do this?

8. In *Yahtzee*, 5 6−sided dice are rolled. How many different outcomes are possible?

9. How many integers solutions are there to $x_1 + x_2 + x_3 + x_4 + x_5 = 60$, where $x_i \geq i$, $i = 1, 2, \ldots, 5$?

10. How many 8−letter strings are able to be made with the 5 vowels in the alphabet if the order is irrelevant?

11. Four (4) letters will be chosen from 26 repetition allowed, and four (4) digits will be chosen from 0 thru 9 with repetition allowed. Assume the order in which the letters and the digits are chosen is irrelevant. We create a code by putting the 4 letters followed by the 4 digits. How many such codes are possible?

12. (Bose-Einstein statistics) Consider a system with B identical bosons (subatomic particles) in a fixed volume and energy level. Suppose the bosons can be in any of g possible states associated with the fixed energy level. In how many ways can these bosons be distributed over the states? In how many ways can these bosons be distributed over the states where at least one state is missing?

II. Probability measures, Axioms and consequences.

The purpose of this section is to introduce probability measures in generality through axioms, which will provide us with properties that *all* probability measures enjoy. More importantly, these properties lead us to develop strategies to compute probabilities.

Motivating the axioms of probability.
When an experiment leads to a sample space Ω with finite and equally likely sample points a natural probability measure to use is the classical probability:

For any event $A \subset \Omega$,

$$P(A) = \frac{|A|}{|\Omega|}.$$

The entire first section of these notes is devoted to methods for computing cardinalities in this model and counting the number of sample points in general, exclusive of the equally likely assumption. We will want to generalize the notion of probability to other models – not just finite equally likely, but also to

- finite, non-equally likely sample spaces
- countable sample spaces like $\mathbb{N} := \{0, 1, 2, 3, \dots\}$ or $\mathbb{Z} := \{\dots, -2, -1, 0, 1, 2, \dots\}$, and
- uncountable sample spaces like the set of real numbers $\mathbb{R}$, $\mathbb{R}^2$ and $\mathbb{R}^n$.

Suppose Ω is a finite sample space. The classical probability has the following 3 essential properties:

1. For any event $A \subseteq \Omega$, $0 \leq P(A) \leq 1$.
This is because $0 \leq |A| \leq |\Omega|$, so $\dfrac{0}{|\Omega|} \leq \dfrac{|A|}{|\Omega|} \leq \dfrac{|\Omega|}{|\Omega|}$.

2. $P(\Omega) = \dfrac{|\Omega|}{|\Omega|} = 1$.

3. If $A_1, A_2, \dots, A_n$ are mutually exclusive ($A_i \cap A_j = \varnothing$ for all $i \neq j$), then
$|A_1 \cup A_2 \cup \cdots \cup A_n| = |A_1| + |A_2| + \cdots |A_n|$, and therefore,
$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$.

These 3 properties of the classical probability measure are properties we would want *all* probability measures to have.

The first one says that the probability is between 0 and 1 (inclusive). For the classical measure, $P(A)$ returns the proportion of all sample points that lie in A. The closer this is to 1 the more likely the event A will occur, the closer to 0, the less likely A will occur.

As for the second property, since Ω has every possible sample point that can occur in the experiment, it is certain that Ω will occur when the experiment is performed.

The last property states that our probability measure should be *additive*: when we union together the sample points from *mutually exclusive events*, the chance the experiment produces a sample point from the union should be the sum of the chances the sample point came from any of these events. Strictly speaking, property 3 above says that P is *finitely* additive since we are only considering finitely many mutually exclusive events. For technical reasons, we will want a similar property to hold for infinite sequences of mutually exclusive events - to have *countable* additivity.

Let P be any probability law, not necessarily the classical one. The following are properties that P is postulated to have.

The 3 Axioms of Probability. (A. Kolmogorov, 1933)

1. ***nonnegativity***:
   For any event A, $0 \leq P(A) \leq 1$.
2. ***normalization***:
   $P(\Omega) = 1$.
3. ***countable additivity***:
   For any sequence of mutually exclusive events $A_1, A_2, A_3, \ldots,$

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots .$$

Remark. ($P(\varnothing) = 0$)
Let's show that these axioms imply $P(\varnothing) = 0$ in general, not just for the classical probability measure. Intuitively, this says the probability no outcome will happen when the experiment is performed is *zero*, and this makes sense because when the experiment is performed we know *some* sample point will occur. Notice $\Omega = \Omega \cup \varnothing \cup \varnothing \cup \varnothing \cdots$ and $\Omega, \varnothing, \varnothing, \varnothing, \ldots$ is a sequence of mutually exclusive events. P is countably additive, so

$$\underbrace{P(\Omega)}_{=1} = \underbrace{P(\Omega)}_{=1} + \underbrace{P(\varnothing)}_{=c} + \underbrace{P(\varnothing)}_{=c} + \underbrace{P(\varnothing)}_{=c} + \cdots .$$

By nonnegativity, $P(\varnothing) \geq 0$, and this equation balances exactly when $P(\varnothing) = c = 0$.

Remark. (countable additivity implies *finite* additivity)
Let $A_1, A_2, \ldots, A_n$ be a *finite* collection of mutually exclusive events. Then

$$A_1, A_2, \ldots, A_n, \varnothing, \varnothing, \varnothing, \ldots$$

is a *sequence* of mutually exclusive events and, using a parallel argument, countable additivity implies

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) &= P(A_1 \cup A_2 \cup \cdots \cup A_n \cup \varnothing \cup \varnothing \cup \cdots) \\ &= P(A_1) + P(A_2) + \cdots + P(A_n) + \underbrace{P(\varnothing)}_{=0} + \underbrace{P(\varnothing)}_{=0} + \cdots \\ &= P(A_1) + P(A_2) + \cdots + P(A_n). \end{aligned}$$

So, a countably additive probability measure is also finitely additive. The converse is *not* true! We will want our probability measures to be countably additive in this course because it will be necessary to consider *countable* unions of mutually exclusive events. In the scheme of what we will cover in this course, the distinction between these two types of additivity is *not* important, but becomes important in more advanced treatments of probability. Of course, in finite sample spaces the two concepts are identical.

When our sample space Ω is finite and equally likely, then the Axioms tell us that the classical probability measure *is* the resulting probability measure for this situation. Let's show this now:

For Ω finite, equally likely, the axioms imply P is the classical probability law. Let $|\Omega| < \infty$, say, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$, and suppose each $\omega \in \Omega$ is equally likely, i.e., there is a constant c such that $P(\{\omega\}) = c$ for all $\omega \in \Omega$. Then,

$$\Omega = \bigcup_{i=1}^{N} \{\omega_i\} = \{\omega_1\} \cup \{\omega_2\} \cup \cdots \cup \{\omega_N\},$$

and because the ω_i's are distinct, the ***singleton sets*** $\{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_N\}$ are mutually exclusive. So, by normalization and *finite additivity*,

$$1 = P(\Omega) = \sum_{i=1}^{N} P(\{\omega_i\}) = \sum_{i=1}^{N} c = cN \implies c = P(\{\omega\}) = \frac{1}{N} \quad \text{for each } \omega \in \Omega.$$

Now, in a similar manner, we can write any event A as a mutually exclusive union of its members: say, $A = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_k}\} = \{\omega_{i_1}\} \cup \{\omega_{i_2}\} \cup \cdots \cup \{\omega_{i_k}\}$. It will then follow

$$P(A) = \underbrace{P(\{\omega_{i_1}\}) + P(\{\omega_{i_2}\}) + \cdots + P(\{\omega_{i_k}\})}_{k \text{ copies of } \frac{1}{N}} = \frac{k}{N} = \frac{|A|}{|\Omega|}.$$

This is the classical probability law! □

We now generalize the classical probability measure:

Discrete probability measures.
Let Ω be finite or countably infinite sample space, $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$. Suppose we have nonnegative numbers $p_1, p_2, p_3, \ldots$ – called ***probability masses*** – such that $\sum_i p_i = 1$, and we assign

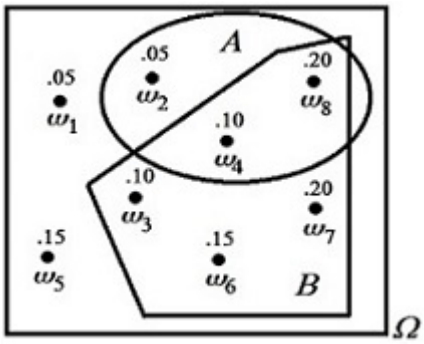
$$P(\{\omega_i\}) = p_i \quad \text{for each } i = 1, 2, 3, \ldots.$$

Then, given an event A, write it as $A = \{\omega_{i_1}\} \cup \{\omega_{i_2}\} \cup \{\omega_{i_3}\} \cdots$ – a mutually exclusive union of its members. By countable additivity,

$$P(A) = P(\{\omega_{i_1}\}) + P(\{\omega_{i_2}\}) + P(\{\omega_{i_3}\}) + \cdots = \sum_{k:\omega_{i_k} \in A} p_{i_k},$$

i.e., to compute the probability of A we just sum the probability masses at all those ω that belong to the event A. This defines the ***discrete probability measure***.

Basic example.
Using the probability masses given in the figure to the right,
Compute
$P(A),\ P(B),\ P(A \cap B),$
$P(A \cup B),\ P(A^c \cap B),\ P(A \cap B^c).$



SOLUTION:
$P(A) = .05 + .10 + .20 = .35.$
$P(B) = .10 + .10 + .15 + .20 + .20 = .75.$
$P(A \cap B) = .10 + .20 = .30.$
$P(A \cup B) = 1 - (.05 + .15) = .80.$
$P(A^c \cap B) = .10 + .15 + .20 = .45.$
$P(A \cap B^c) = .05.$

Example.
An experiment has a sample space $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. Experience dictates that ω_2 and ω_3 are equally likely, but are each twice as likely as ω_1; and, ω_4 is 4 times as likely as ω_1. Construct the discrete probability masses associated with each sample point.

SOLUTION:
If we set $P(\{\omega_1\}) = x$, then we are told $P(\{\omega_2\}) = P(\{\omega_3\}) = 2x$ and $P(\{\omega_4\}) = 4x$. The normalization axiom says

$$P(\Omega) = x + 2x + 2x + 4x = 9x = 1 \implies x = \frac{1}{9}.$$

Therefore,

$$P(\{\omega_1\}) = \frac{1}{9} =: p_1,\ P(\{\omega_2\}) = \frac{2}{9} =: p_2,\ P(\{\omega_3\}) = \frac{2}{9} =: p_3, \text{ and } P(\{\omega_4\}) = \frac{4}{9} =: p_4.$$

Example. (the geometric($\frac{1}{2}$) distribution - modeling probability masses)
Toss a fair coin until the occurrence of the first head. The sample space $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4 \dots\}$ is countably infinite, and we can think of ω_i as the outcome that the first head occurs on the ith toss:

$\omega_1 = h$
$\omega_2 = th$
$\omega_3 = tth$
$\omega_4 = ttth$
$\vdots$
$\omega_i = \underbrace{tt\cdots t}_{i-1 \text{ tails}}h$

If $p_i = P(\{\omega_i\}) = P(\{tt\cdots th\})$, then what would be a reasonable model for the p_i ?
Use this model to compute the probability the first head occurs *before* the 4th toss. How about an *even* toss?

The coin is fair, so I would expect $p_1 = P(\{\omega_1\}) = \dfrac{1}{2}$ – half the time the first toss is a head, half the time a tail.

How about $P(\{\omega_2\}) = P(\{th\})$?

If we think of tossing a fair coin repeatedly, then eventually we will surpass the second toss in the experiment. The event $\{th\}$ can be thought of as tossing a coin just twice and getting a tail followed by a head. This is only one sample point of $2^2 = 4$ equally likely possibilities, hh, ht, th, tt, that can result when tossing the coin twice. So, I would expect $p_2 = P(\{\omega_2\}) = \dfrac{1}{2^2} = \left(\dfrac{1}{2}\right)^2$.

In fact, $p_i = P(\{tt\cdots th\}) = \left(\dfrac{1}{2}\right)^i$.

The event A that the first head occurs before the 4th toss is $\{\omega_1, \omega_2, \omega_3\}$. So,

$$P(A) = P(\{\omega_1\}) + P(\{\omega_2\}) + P(\{\omega_3\}) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}.$$

For the event B that the first head occurs on an even toss, i.e., $B = \{\omega_2, \omega_4, \omega_6, \omega_8, \dots\}$, and

$$\begin{aligned} P(B) &= P(\{\omega_2\}) + P(\{\omega_4\}) + P(\{\omega_6\}) + P(\{\omega_8\}) + \cdots \\ &= \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^8 + \cdots \qquad (\text{geomtric ratio} = \tfrac{1}{4}) \\ &= \frac{\dfrac{1}{4}}{1 - \dfrac{1}{4}} \\ &= \frac{1}{3}. \end{aligned}$$

In the calculation of $P(B)$ above I used following:

Calculus fact: Sums of a geometric series.
For any constants A and r, the sequence $A, Ar, Ar^2, Ar^3, \dots$ is called a ***geometric progression***. The value r is called the ***geometric ratio***. When $|r| < 1$ the ***sum of the geometric series*** is

$$\sum_{k=m}^{\infty} Ar^k = \frac{Ar^m}{1 - r}.$$

Advice:
Memorize this formula for the sum of a geometric series. It will be used several times in this course. A good mnemonic for remembering the formula for a geometric series is "it's the first term in the series divided by 1 minus the geometric ratio".

Consequences of the Axioms: Properties of probability measures.

The axioms will imply some properties that all probability measure will possess.
I name a few:

- The Complementary Rule.
- Monotonicity.
- Subadditivity.
- Inclusion-exclusion rules.

and others.

The Complementary Rule.

For any event A,

$$P(A) = 1 - P(A^c).$$

Proof.
A and A^c are mutually exclusive and $A \cup A^c = \Omega$. So, by finite additivity,

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c).$$

□

Advice:
Computing the probability of the complement event may be much easier than computing the probability of the event directly. You should keep this in mind!

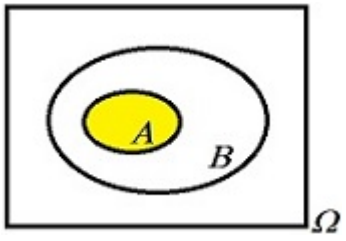
Monotonicity.

If A and B are events, and $A \subseteq B$, then

$$P(A) \leq P(B).$$

Proof.
Notice $B = A \cup (A^c \cap B)$ and the events A and $A^c \cap B$ are mutually exclusive. Therefore, by finite additivity and nonnegativity,

$$P(B) = P(A) + \underbrace{P(A^c \cap B)}_{\substack{\geq 0 \text{ by} \\ \text{nonnegativity}}} \geq P(A) + 0 = P(A).$$



□

Remark.
Basically monotonicity says if you add sample points to an event, the probability cannot become smaller.

Subadditivity. (also called Boole's inequality)
For *any* events $A_1, A_2, \ldots, A_n$,

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n),$$

or, succinctly,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

Of course, if the events were mutually exclusive, then we have equality, but in general, Boole's inequality provides a very crude upper bound for the probability of the union.

Proof.
I will proceed by induction. When $n = 1$ there's nothing to prove, we have a tautology. So, consider the case $n = 2$. For any events A_1 and A_2, $A_1 \cup A_2 = A_1 \cup (A_1^c \cap A_2)$, where A_1 and $A_1^c \cap A_2$ are mutually exclusive (see the picture).

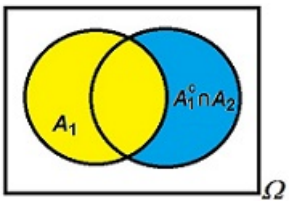


Figure. A_1 (yellow) and $A_1^c \cap A_2$ (blue) are mutually exclusive.

Therefore, $P(A_1 \cup A_2) = P(A_1) + P(A_1^c \cap A_2)$. But, since $A_1^c \cap A_2 \subseteq A_2$, by monotonicity, $P(A_1^c \cap A_2) \leq P(A_2)$. Thus,

$$P(A_1 \cup A_2) = P(A_1) + P(A_1^c \cap A_2) \leq P(A_1) + P(A_2),$$

and the statement is true for $n = 2$ events.

Suppose, for some *fixed* $k \geq 2$, the statement is true for any events $A_1, A_2, \ldots, A_k$; namely

$$P(A_1 \cup A_2 \cup \cdots \cup A_k) \leq \sum_{i=1}^{k} P(A_i).$$

Let $A_1, A_2, \ldots, A_k, A_{k+1}$ be any $k + 1$ events.

$$\begin{aligned}
P\left(\bigcup_{i=1}^{k+1} A_i\right) &= P\left([A_1 \cup A_2 \cup \cdots \cup A_k] \cup A_{k+1}\right) \\
&\leq P\left(A_1 \cup A_2 \cup \cdots \cup A_k\right) + P\left(A_{k+1}\right) \qquad (n = 2 \text{ case}) \\
&\leq \sum_{i=1}^{k} P(A_i) + P(A_{k+1}) \qquad \text{(induction hypothesis)} \\
&= \sum_{i=1}^{k+1} P(A_i),
\end{aligned}$$

and this completes the induction. □

Two useful results when working with unions of events are the inclusion-exclusion rules and DeMorgan's laws. We discuss each in turn. The results that follow are true for any events $A_1, A_2, \ldots, A_n$ unless noted otherwise.

The inclusion-exclusion rules.
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ -P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

$$P(A \cup B \cup C \cup D) = P(A) + P(B) + P(C) + P(D) \\ -P(A \cap B) - P(A \cap C) - P(A \cap D) - P(B \cap C) - P(B \cap D) - P(C \cap D) \\ +P(A \cap B \cap C) + P(A \cap B \cap D) + P(A \cap C \cap D) + P(B \cap C \cap D) \\ -P(A \cap B \cap C \cap D).$$

In general,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \underbrace{\sum_{i=1}^{n} P(A_i)}_{\binom{n}{1} \text{ terms}} - \underbrace{\sum\sum_{1 \le i_1 < i_2 \le n} P(A_{i_1} \cap A_{i_2})}_{\binom{n}{2} \text{ terms}} + \underbrace{\sum\sum\sum_{1 \le i_1 < i_2 < i_3 \le n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3})}_{\binom{n}{3} \text{ terms}} - + \cdots$$

Remark.
The inclusion exclusion rules are very useful to compute the probability of a union of events when probabilities of intersections of the events are easier to compute. In fact, in many applications of this result it will turn out that the probabilities of the intersections of a fixed number of events will all be the same value, making it even easier to use than may first appear.

Advice:
The rules are not very hard to remember if you look at the patterns. We *add* the probabilities of the all n single sets, then *subtract* the probabilities of all the 2-way intersections, than *add* the probabilities of all the 3-way intersections, and so forth, alternating additions and subtractions until we get to the n-way intersection.

The Boole-Bonferroni inequalities.

$$P\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} P(A_i)$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \ge \sum_{i=1}^{n} P(A_i) - \sum\sum_{1 \le i_1 < i_2 \le n} P(A_{i_1} \cap A_{i_2})$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} P(A_i) - \sum\sum_{1 \le i_1 < i_2 \le n} P(A_{i_1} \cap A_{i_2}) + \sum\sum\sum_{1 \le i_1 < i_2 < i_3 \le n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3})$$

$$\vdots$$

Let's prove the inclusion-exclusion rule for 2 sets, then show how the rule for more sets can follow inductively without giving a formal proof.

Proof of the inclusion exclusion rule for 2 sets:
For $n = 2$ events, A_1, A_2, we can write $A_1 \cup A_2 = A_1 \cup (A_1^c \cap A_2)$ where A_1 and $A_1^c \cap A_2$ are mutually exclusive. Therefore, $P(A_1 \cup A_2) = P(A_1) + P(A_1^c \cap A_2)$.

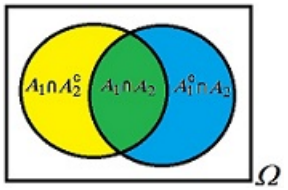


Figure. Decomposition of $A_1 \cup A_2$ into mutually exclusive pieces.

However, $A_2 = (A_1 \cap A_2) \cup (A_1^c \cap A_2)$ and, again, $A_1 \cap A_2$ and $A_1^c \cap A_2$ are mutually exclusive, so $P(A_2) = P(A_1 \cap A_2) + P(A_1^c \cap A_2)$ implying $P(A_1^c \cap A_2) = P(A_2) - P(A_1 \cap A_2)$. Substituting this into the equation above the picture we get

$$P(A_1 \cup A_2) = P(A_1) + \underbrace{P(A_1^c \cap A_2)}_{=P(A_2)-P(A_1 \cap A_2)} = P(A_1) + P(A_2) - P(A_1 \cap A_2),$$

which shows the statement is true in the case of $n = 2$ events. □

The inclusion-exclusion rule for 2 sets implies the rule for 3 sets:
Now consider any 3 events A_1, A_2 and A_3.

$$\begin{aligned}
P(A_1 \cup A_2 \cup A_3) &= P([A_1 \cup A_2] \cup A_3) \\
&= P(A_1 \cup A_2) + P(A_3) - P([A_1 \cup A_2] \cap A_3) \qquad \text{(by the rule for 2 sets)} \\
&= P(A_1) + P(A_2) - P(A_1 \cap A_2) + P(A_3) - P([A_1 \cap A_3] \cup [A_2 \cap A_3]) \\
&= P(A_1) + P(A_2) - P(A_1 \cap A_2) + P(A_3) \\
&\quad - \Big(P(A_1 \cap A_3) + P(A_2 \cap A_3) - P(\underbrace{[A_1 \cap A_3] \cap [A_2 \cap A_3]}_{=A_1 \cap A_2 \cap A_3}) \Big) \\
&= P(A_1) + P(A_2) + P(A_3) \\
&\quad - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\
&\quad + P(A_1 \cap A_2 \cap A_3).
\end{aligned}$$

Exercise for the student.
Now that we showed the inclusion-exclusion rule for 2 sets and for 3 sets, show how the rule for 4 sets follows from these lower order rules.

DeMorgan's laws.

For any events A and B,

$$\big(A \cup B\big)^c = A^c \cap B^c \quad \text{and} \quad \big(A \cap B\big)^c = A^c \cup B^c$$

and, more generally, for *any* number of events

$$\Big(\bigcup_i A_i\Big)^c = \bigcap A_i^c \quad \text{and} \quad \Big(\bigcap_i A_i\Big)^c = \bigcup_i A_i$$

Proof.
$\omega \in (A \cup B)^c$ iff $\omega \notin A \cup B$ iff ω is *not* in at least one of A or B iff ω is *not* in A and ω is *not* in B iff $\omega \in A^c \cap B^c$. Therefore, $(A \cup B)^c = A^c \cap B^c$.

Useful fact.
By the complementary rule,

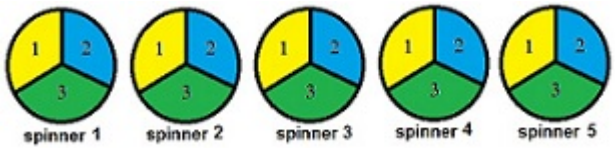
$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= 1 - P\left(\Big(\bigcup_{i=1}^n A_i\Big)^c\right) \\ &= 1 - P\left(\bigcap_{i=1}^n A_i^c\right) \qquad \text{(by DeMorgan's law)}, \end{aligned}$$

and, replacing A_i with A_i^c everywhere,

$$P\left(\bigcup_{i=1}^n A_i^c\right) = 1 - P\left(\bigcap_{i=1}^n A_i\right).$$

Remark.
This fact is especially useful to compute the probability of a union of events when computing the probability of the intersection of the *complements* of the events is easier; in particular, such will be the case when the events $A_1, A_2, \ldots, A_n$ happen to be *independent*, which we will discuss shortly (see page 90).

Example and discussion.
We have 5 spinners labeled 1 thru 5 each having 3 equally likely regions numbered 1,2,3 as in this picture:



Here $\Omega = \{(x_1, x_2, x_3, x_4, x_5) : x_i \in \{1, 2, 3\},\ i = 1, 2, 3, 4, 5\}$ is finite and equally likely. We spin each spinner. What's the probability that we are *void* in at least one number?

SOLUTION:
Being void in a number means that a value (1, 2, or 3) is missing in the outcome. Here are some sample points (ordered 5-tuples):

$$\underbrace{(1,3,3,1,1)}_{\text{void in 2's}} \quad , \quad \underbrace{(2,2,3,2,2)}_{\text{void in 1's}} \quad , \quad \underbrace{(3,3,3,3,3)}_{\text{void in 1 and 2's}} \quad \text{and} \quad \underbrace{(1,2,1,2,2)}_{\text{void in 3's}}.$$

If we define V_i to be the event that we are void in the number i, then we are interested in computing the probability of $V_1 \cup V_2 \cup V_3$ – this is the event that *at least one* of the events V_1, V_2 or V_3 happens which is exactly what we want! I claim this is a situation where computing the probabilities of the intersections of these events is easier (than computing the intersection of their complements). Therefore, I will employ the inclusion-exclusion rule for 3 sets here (see the second formula at the top of page 64):

$$\begin{aligned} P(V_1 \cup V_2 \cup V_3) \;=\; & P(V_1) + P(V_2) + P(V_3) \\ & - P(V_1 \cap V_2) - P(V_1 \cap V_3) - P(V_2 \cap V_3) \\ & + P(V_1 \cap V_2 \cap V_3). \end{aligned}$$

The event V_1 has all the ordered 5-tuples that are missing the number 1, i.e., can only have entries from the set $\{2, 3\}$. This is just sampling with replacement 5 times from a set of size 2: $|V_1| = 2^5$. Likewise, $|V_2| = |V_3| = 2^5$. The event $V_1 \cap V_2$ are the ordered 5-tuples that are missing *both* the numbers 1 and 2, i.e., can only have the one entry $\{3\}$. So, $V_1 \cap V_2 = \{(3, 3, 3, 3, 3)\}$ and $|V_1 \cap V_2| = 1$. Similarly, $|V_1 \cap V_3| = |V_2 \cap V_3| = 1$. Lastly, $V_1 \cap V_2 \cap V_3 = \varnothing$, so $|V_1 \cap V_2 \cap V_3| = 0$.

Putting these calculations into the inclusion-exclusion rule above and noting that $|\Omega| = 3^5$ we obtain

$$P(V_1 \cup V_2 \cup V_3) = 3 \cdot \frac{2^5}{3^5} - 3 \cdot \frac{1}{3^5} + \frac{0}{3^5} = \frac{31}{81}.$$

Further discussion on the spinners example.
What if we had attempted to use the formula $P(V_1 \cup V_2 \cup V_3) = 1 - P(V_1^c \cap V_2^c \cap V_3^c)$? $V_1^c \cap V_2^c \cap V_3^c$ is the event that we are not void in any of the numbers 1,2 or 3; i.e., there is at least one of each number. There are two cases to consider:

case 1: 3 of one number, one each of other two.

case 2: 2 each of two numbers, one of the remaining.

In case 1: there are $\binom{3}{1}$ ways to select the number that will be tripled, and once this is done, there are $\binom{5}{3}$ of the spinners that this number can appear on, once this is done, there are 2! ways the remaining two numbers can appear on the remaining two spinners. The probability is $\dfrac{\binom{3}{1}\binom{5}{3}\cdot 2!}{3^5}$.

In case 2: there are $\binom{3}{2}$ ways to select the 2 numbers for the two numbers since the doubles are indistinguishable, once this is done there are $\binom{5}{2,2,1}$ ways to position to numbers in the 5-tuple. The probability is $\dfrac{\binom{3}{2}\binom{5}{2,2,1}}{3^5}$.

The result is

$$P(V_1^c \cap V_2^c \cap V_3^c) = \frac{3\cdot\binom{5}{3}\cdot 2! + 3\binom{5}{2,2,1}}{3^5} = \frac{150}{3^5} = \frac{50}{81}.$$

Consequently,

$$P(V_1 \cup V_2 \cup V_3) = 1 - \frac{50}{81} = \frac{31}{81}.$$

You might think that, comparatively, these two computations were about the same degree of difficulty. But, what if there had been 20 spinners instead of 5 and even more equally likely regions than 3? The calculation using the inclusion exclusion rule is no more difficult than the one with 5 spinners; however, the calculation involving the complement event is nightmarish! Of course, if we are very careful one could use multiset counting to enumerate the cases, but you can see the immediate challenges with this approach.

Example. ("Men with hats" problem)
Suppose n men wearing hats put their hats in a room. When they leave each man grabs a hat uniformly at random the hats remaining. What's the probability that no man selects his own hat?

For example, when $n = 3$, $\Omega = \{(1,2,3),(1,3,2),(2,1,3),(2,3,1),(3,1,2),(3,2,1)\}$, where, for instance, $(2,1,3)$ means person 1 selects person 2's hat, person 2 selects person 1's hat, person 3 selects person 3's hat. Out of the $|\Omega| = 3!$ possibilities, only 2, namely $(2,3,1)$ and $(3,1,2)$, have no man selecting their own hat – so, the probability is $\frac{2}{3!} = \frac{1}{3}$.

What do we do with the case of general n?

Let A be the event that no man selects their own hat. Then A^c is the event that at least one man selects their own hat; i.e., $A^c = \bigcup_{i=1}^n M_i$, where M_i is the event that the ith man selects his own hat. The sample space Ω of this experiment is the set of all $n!$ orderings of the integers 1 thru n, and it is important to recall the *exchangeability property* this sample space exhibits.

$P(M_i) = \dfrac{1}{n}$ for $1 \le i \le n$.
$P(M_{i_1} \cap M_{i_2}) = \dfrac{1}{(n)_2}$ for $1 \le i_1 < i_2 \le n$.
$P(M_{i_1} \cap M_{i_2} \cap M_{i_3}) = \dfrac{1}{(n)_3}$ for $1 \le i_1 < i_2 < i_3 \le n$.
In general, $P(M_{i_1} \cap M_{i_2} \cap \cdots \cap M_{i_k}) = \dfrac{1}{(n)_k}$ for $1 \le i_1 < i_2 < \cdots < i_k \le n$.

By the inclusion-exclusion rule for n events,

$$\begin{aligned} P\left(\bigcup_{i=1}^n M_i\right) &= \sum P(M_i) - \sum\sum P(M_{i_1} \cap M_{i_2}) + \sum\sum\sum P(M_{i_1} \cap M_{i_2} \cap M_{i_3}) - + \cdots \\ &= \sum \frac{1}{n} - \sum\sum \frac{1}{(n)_2} + \sum\sum\sum \frac{1}{(n)_3} - \sum\sum\sum\sum \frac{1}{(n)_4} + - \cdots \\ &= n \cdot \frac{1}{n} - \binom{n}{2}\frac{1}{(n)_2} + \binom{n}{3}\frac{1}{(n)_3} - \binom{n}{4}\frac{1}{(n)_4} + - \cdots + (-1)^{n+1}\frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \frac{1}{5!} - + \cdots + (-1)^{n+1}\frac{1}{n!} = P(A^c). \end{aligned}$$

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} - + \cdots + \frac{(-1)^n}{n!} \approx e^{-1} = 0.367879\ldots,$$

where we recognize the MacLaurin series expansion of the exponential function:

Calculus fact: MacLaurin series for e^u.
For any (real) u,

$$e^u = \sum_{k=0}^{\infty} \frac{u^k}{k!} = 1 + u + \frac{u^2}{2!} + \frac{u^3}{3!} + \frac{u^4}{4!} + \cdots .$$

Continuity of probability measures.

Let $E_1, E_2, E_3, \ldots$ be a ***nested increasing sequence of events***: this means, for every n, $E_n \subseteq E_{n+1}$,

$$E_1 \subseteq E_2 \subseteq E_3 \subseteq \cdots \subseteq E_n \subseteq E_{n+1} \subseteq \cdots .$$

Each event in this sequence is contained in the next, therefore, for every n,

$$\bigcup_{i=1}^{n} E_i = E_n.$$

So, if we set $E = \bigcup_{i=1}^{\infty} E_i$, then $\bigcup_{i=1}^{n} E_i \to \bigcup_{i=1}^{\infty} E_i =: E$ as $n \to \infty$, which we will abbreviate as $E_n \uparrow E$ as $n \to \infty$. The figure below show a nested increasing sequence of events $E_1 \subseteq E_2 \subseteq \cdots$ and its decomposition into mutually exclusive events $E_1, E_2 - E_1, E_3 - E_2, \ldots$.

Here's a similar situation with a ***nested decreasing sequence of events***:

$$F_1 \supseteq F_2 \supseteq F_3 \supseteq \cdots \supseteq F_n \supseteq F_{n+1} \supseteq \cdots .$$

Each event in this sequence contains the next, therefore, for every n,

$$\bigcap_{i=1}^{n} F_i = F_n.$$

So, if we set $F = \bigcap_{i=1}^{\infty} F_i$, then $\bigcap_{i=1}^{n} F_i \to \bigcap_{i=1}^{\infty} F_i =: F$ as $n \to \infty$, which we will abbreviate as $F_n \downarrow F$ as $n \to \infty$.

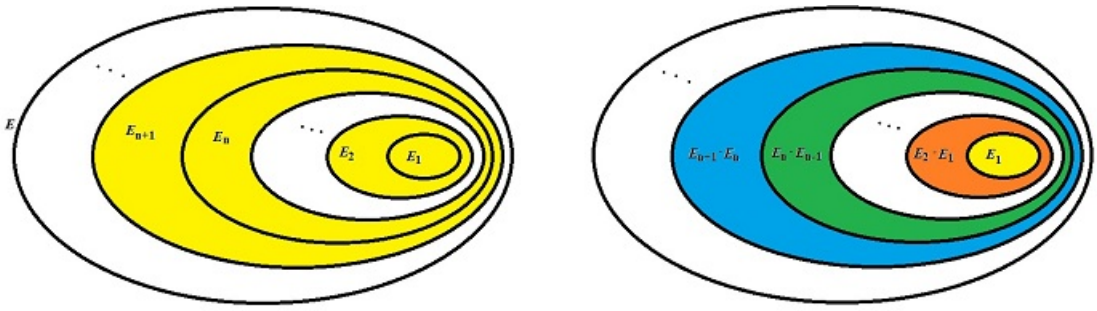


Figure. On left, nested events $E_1 \subseteq E_2 \subseteq \cdots$; on right, decomposition into mutually exclusive pieces.

Here's the result I'd like to show:

Continuity of probability.
If $E_n \uparrow E$ as $n \to \infty$, then $P(E_n) \to P(E)$ as $n \to \infty$.
If $F_n \downarrow F$ as $n \to \infty$, then $P(F_n) \to P(F)$ as $n \to \infty$.

Proof.
Suppose $E_n \uparrow E$. Then, we can write as a union of mutually exclusive events:

$$E = E_1 \cup \bigcup_{i=1}^{\infty} \Big(E_{i+1} - E_i \Big).$$

Therefore,

$$\begin{aligned}
P(E) &= P(E_1) + \sum_{i=1}^{\infty} P(E_{i+1} - E_i) \quad \text{(by countable additivity)} \\
&= P(E_1) + \sum_{i=1}^{\infty} \Big(P(E_{i+1}) - P(E_i) \Big) \quad (E_i \subseteq E_{i+1}) \\
&= \lim_{n\to\infty} \left(P(E_1) + \sum_{i=1}^{n} \Big(P(E_{i+1}) - P(E_i) \Big) \right) \quad \text{(series is a limit of its partial sums)} \\
&= \lim_{n\to\infty} P(E_n). \qquad \text{(the partial sums are telescoping)}
\end{aligned}$$

Now, suppose $F_n \downarrow F$ as $n \to \infty$. Then $F_n^c \uparrow F^c$. By what we just proved:

$$\lim_{n\to\infty} P(F_n^c) = P(F^c) = 1 - P(F).$$

But, $\lim_{n\to\infty} P(F_n^c) = \lim_{n\to\infty} \Big(1 - P(F_n) \Big) = 1 - \lim_{n\to\infty} P(F_n)$. Putting this together with the above it follows that

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

which completes the proof. □

Conditional probability.

Sometimes partial information about the outcome of an experiment comes available, and this extra information may change the probability of events. For example, two fair 6-sided dice are rolled, the probability the sum is 10 is $\frac{3}{36}$ because out of the 36 equally likely sample points, 3 of them sum to 10. But, suppose when the dice are thrown someone saw the number 1 was on at least one of the dice. Given this information, we see it is now impossible for the sum to be 10, so indeed, this partial information changed the probability the sum is 10. We will say the conditional probability the sum is 10 given a 1 appeared is *zero*.

Carrying this one step further, the probability that the sum is 7 is $\frac{1}{6}$ without any additional information. Now consider the question:

If a 1 appears, what's the probability the sum is 7 ?

First of all, this question only cares about the probability that the sum is 7 under the specific condition, and thus, we are trying to compute a conditional probability. The *given information* in the condition is the *event* 'a 1 appears'. So, we need to compute the probability that the sum of the dice will be 7 given that the event 'a 1 appears' has occurred.

Let's look at the sample space (xy means x rolled first, y rolled second):

$$\begin{aligned}\Omega = \{&11, \ 12, \ 13, \ 14, \ 15, \ 16,\\ &21, \ 22, \ 23, \ 24, \ 25, \ 26,\\ &31, \ 32, \ 33, \ 34, \ 35, \ 36,\\ &41, \ 42, \ 43, \ 44, \ 45, \ 46,\\ &51, \ 52, \ 53, \ 54, \ 55, \ 56,\\ &61, \ 62, \ 63, \ 64, \ 65, \ 66\}\end{aligned}$$

Under the given condition 'a 1 appears', the sample space Ω as written has too many sample points – we can ignore, for instance, 66 since a 1 doesn't appear in it. The given information ***reduces*** the sample space to just those sample points in the given event. Here's the reduced sample space:

$$\begin{aligned}\Omega = \{&11, \ 12, \ 13, \ 14, \ 15, \ 16,\\ &21,\\ &31,\\ &41,\\ &51,\\ &61, \qquad\qquad\qquad\quad \}\end{aligned}$$

I'll emphasize that we cannot tell if the die we saw with the number 1 on it was rolled first or was rolled second, which hopefully explains why we needed to include all 11 sample points that have at least one 1 in them. The original sample space was equally likely, so these 11 sample points are also equally likely, but only 2 of these sum to 7, namely, 61 and 16; so the conditional probability the sum is 7 given a 1 appears is $\frac{2}{11}$.

To follow-up a bit..., suppose the two dice were different colors – say, one is red, the other is green – and, we are interested in computing the (conditional) probability that the sum is 7 given the red die shows a 1. Now, the original sample space of 36 equally likely sample points is reduced to only these 6 sample points (assuming xy means x on the red, y on the green die):

$$\Omega = \{11,\ 12,\ 13,\ 14,\ 15,\ 16\}$$

Among these 6 equally likely sample points only 1 sums to 7, so the conditional probability is $\frac{1}{6}$, i.e., the information that the red die shows a 1 doesn't alter the unconditional probability the sum is 7.

Notation:
We write the (conditional) probability of A given B as $P(A|B)$. It represents the probability of A under the condition the event B occurs.

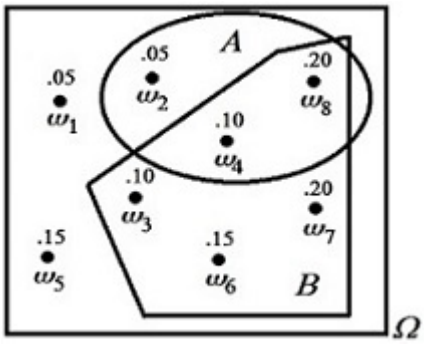
The conditional probability formula.
For any event A, when $P(B) > 0$, we can compute

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

Advice:
Memorize this formula. Especially remember that $P(B)$ needs to be positive to use it.

The motivation behind this formula is that if we are thinking that B is the new sample space, then the sample points within it should occur in the same relative proportions. Thus, we should normalize the probabilities of each sample point within B by $P(B)$ (of course, we'd need $P(B) > 0$ to divide by it!). After doing this all the probabilities in B will sum to 1 making it a valid sample space. If we're given B occurs, then only the portion of A that belongs to B will matter (i.e., $A \cap B$); the probability of A in this new sample space is really the ratio of the unconditional probabilities given in the formula. I'll illustrate what's going on in the next example.

Example.
With the discrete probability space pictured to the right, compute $P(A|B)$.

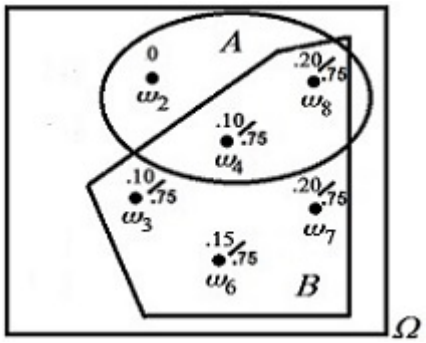


SOLUTION:
The event $A = \{\omega_2, \omega_4, \omega_8\}$, and $B = \{\omega_3, \omega_4, \omega_6, \omega_7, \omega_8\}$.
Using the conditional probability formula:

$P(A \cap B) = P(\{\omega_4\}) + P(\{\omega_8\}) = .10 + .20 = .30.$

Similarly, $P(B) = .75$. Therefore,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.30}{.75} = \frac{6}{15}.$$



Alternatively, knowing B is the "new" sample space, the sample points within B occur in the same relative proportions. Therefore, the probabilities of each sample point in B should be normalized by $P(B) = .75$. (see picture to right: B is now a sample space!) only the sample points in $A \cap B$ are now possible, as it is impossible for ω_2 to occur given B. The remaining sample points in $A \cap B$ should occur in the same proportions within B. Since $A \cap B = \{\omega_4, \omega_8\}$, $P(A|B) = \frac{.10}{.75} + \frac{.20}{.75} = \frac{.30}{.75}$ as before.

Exercise for students.
Fix an event B in a sample space Ω with $P(B) > 0$. We *define* the set function

$$P_B(A) := \frac{P(A \cap B)}{P(B)} \quad \text{for each event } A \subseteq \Omega.$$

Show that P_B satisfies the 3 axioms of a probability and *is*, therefore, a probability measure. Moral of the story: For *fixed* B with $P(B) > 0$, the conditional probability $P(\cdot|B)$ is a probability measure (just with B as its sample space).

Example. (continued from the geometric($\frac{1}{2}$) example from page 60)
You toss the fair coin repeatedly. If the first head occurs on an even-numbered toss, what's the probability it was the second toss? Also answer: If the first head occurs on an odd-numbered toss, what's the probability it was the first toss?

SOLUTION:
Define B to be the event the first head occurs on an even-numbered toss, and let A_i be the event that the first head occurs on the ith toss. We are interested in $P(A_2|B)$. On page 60, we learned that $P(B) = \frac{1}{3} > 0$. Now, if we compute $P(A_2 \cap B)$ then we can apply the conditional probability formula. But, in this problem, $A_2 \subseteq B$ since the second toss is also an even-numbered toss; therefore, $A_2 \cap B = A_2$. $P(A_2 \cap B) = P(A_2) = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$. Finally,

$$P(A_2|B) = \frac{\frac{1}{4}}{\frac{1}{3}} = \frac{3}{4}.$$

Since the first head occurs on either an even-numbered toss or an odd-numbered toss and not both, B^c is the event the first head occurs on an odd-numbered toss. By the complementary rule, $P(B^c) = 1 - P(B) = \frac{2}{3}$. We are now interested in $P(A_1|B^c)$. Again, $A_1 \subseteq B^c$, so $P(A_1 \cap B^c) = P(A_1) = \frac{1}{2}$. Therefore, the probability the first toss is a head *given* the first head occurs on an odd-numbered toss is

$$P(A_1|B^c) = \frac{\frac{1}{2}}{\frac{2}{3}} = \frac{3}{4}.$$

Example.
I deal you a 5-card hand from a standard deck of 52 cards. What's the probability they are all hearts given they are all red?

SOLUTION:
Let R be the event all cards are red, and H the event that all cards are hearts. We are interested in $P(H|R)$. Sometimes, when working with sample spaces that were originally equally likely (equally likely before the given information that all cards are red), it's often better to work with the reduced sample space, especially when the event of interest is a subset of the given event (as it is here: all hearts is a subset of all red):

$$P(H|R) = \frac{P(H \cap R)}{P(R)} = \frac{P(H)}{P(R)} = \frac{\frac{|H|}{|\Omega|}}{\frac{|R|}{|\Omega|}} = \frac{|H|}{|R|}.$$

In this case, $|R| = \binom{26}{5}$ and $|H| = \binom{13}{5}$. Therefore,

$$P(H|R) = \frac{\binom{13}{5}}{\binom{26}{5}} \approx 0.02.$$

By the way, the unconditional probability $P(H) = \dfrac{|H|}{|\Omega|} = \dfrac{\binom{13}{5}}{\binom{52}{5}} \approx 0.0005$, quite a difference!

Example. (continued from the 5 spinners problem on page 67)
You spin the 5 spinners. Compute the probability that the number 3 is missing given you are void in a number.

SOLUTION:
Here's a problem where intuition can lead you astray if you are not careful. You *might* think the answer should be $\frac{1}{3}$ because we are equally likely to be missing any of the three numbers if we are told we are void, and, this would be true if the only way we can be void in a number is to be void in exactly one number. However, we can be void in 3's simultaneously with being void in 1's or 2's.

This problem is asking us to compute $P(V_3|V_1 \cup V_2 \cup V_3)$.

$$P(V_3|V_1 \cup V_2 \cup V_3) = \frac{P(V_3 \cap (V_1 \cup V_2 \cup V_3))}{P(V_1 \cup V_2 \cup V_3)} = \frac{P(V_3)}{P(V_1 \cup V_2 \cup V_3)} = \frac{\frac{2^5}{3^5}}{\frac{31}{81}} = \frac{32}{93} \approx 0.344.$$

Example.
We have 30 balls: 16 are red, 8 are green, and 6 are yellow. We line up the balls left to right. Given all the green balls are consecutive, what's the probability there are no two adjacent yellows?

SOLUTION:
Let G be the event that all green balls are consecutive, and let Y be the event that there are no adjacent yellows. I'll work in the equally likely sample space of all anagrams of the 30-letter word with 16 R's, 8 G's and 6 Y's: $|\Omega| = \frac{30!}{16!8!6!}$. G is the event that the anagram has all G's in a row. So, treat the 8 G's as a single super-letter $\tilde{G}$. The resulting word will have 23 letters now: 1 super-G, 16 R's and 6 Y's. The number of anagrams is, therefore, $\frac{23!}{1!16!6!}$

$$P(G) = \frac{\frac{23!}{1!16!6!}}{\frac{30!}{16!8!6!}}.$$

Now let's look at the event $Y \cap G$: keep the Y's separated *and* the G's together.

$$\_\tilde{G}\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_R\_$$

To keep the Y's separated in this word, we need to select 6 of the $17 + 1 = 18$ blanks to put the identical Y's into; and, for each such choice, there are $\frac{17!}{1!16!} = 17$ anagrams that keep the greens together. Therefore,

$$P(Y \cap G) = \frac{\binom{18}{6}17}{\frac{30!}{16!8!6!}}.$$

Therefore, by the conditional probability formula

$$P(Y|G) = \frac{P(Y \cap G)}{P(G)} = \frac{\frac{\binom{18}{6}17}{\frac{30!}{16!8!6!}}}{\frac{\frac{23!}{1!16!6!}}{\frac{30!}{16!8!6!}}} = \frac{\binom{18}{6}17}{\frac{23!}{16!6!}} \approx 0.18.$$

Remark.
Sometimes conditional probabilities are given or are easier to compute/understand in a given situation than some related unconditional probabilities. By rewriting the conditional probability formula as

$$P(A \cap B) = P(A|B)P(B) \qquad \text{(assuming } P(B) > 0\text{)}$$

or as

$$P(A \cap B) = P(B|A)P(A) \qquad \text{(assuming } P(A) > 0\text{)}$$

can be especially useful. These formulas are sometimes called the ***multiplicative rule of conditional probability***.

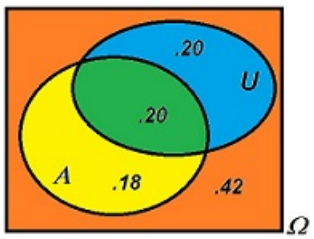
Basic example.
Among left handed people, 50% exhibit a counter-clockwise cowlick (spiral behavior of hair on head). 10% of people are left handed. What proportion of people are both left-handed and exhibit a counter-clockwise cowlick?

SOLUTION:
Let L be the set of lefthanded people and C the event of exhibiting a counter-clockwise cowlick. We are told $P(C|L) = .50$ and $P(L) = .10$. We want $P(C \cap L)$. By the multiplicative rule of conditional probability $P(C \cap L) = P(C|L)P(L) = 0.50 \cdot 0.10 = 0.05$. So, 5% of people are lefthanded and have counter-clockwise cowlicks.

It should be clear that $P(L|C)$ is something entirely different here. This is the proportion of people with counter-clockwise cowlicks that are lefthanded.

Example.
In a typical year 40% of Probability students are underclassmen (Freshmen and Sophomore). Among underclassmen, 50% get an A, while among non-underclassmen 30% get an A. Let U be the set of underclassmen, and let A be the set of students who receive an A in Probability.
(a) Find the proportion of students who receive an A in Probability.
(b) What proportion of students who receive an A in Probability are underclassmen?

SOLUTION:
Sometimes a Venn diagram can help facilitate a solution.



We are told $P(U) = .40$, $P(A|U) = .50$ and $P(A|U^c) = .30$.
The event $A = (A \cap U) \cup (A \cap U^c)$.
By the multiplicative rule of conditional probability
$P(A \cap U) = P(A|U)P(U) = .50 \cdot .40 = .20$ (green region).
$P(A \cap U^c) = P(A|U^c)P(U^c) = .30 \cdot (1 - .40) = .18$ (yellow region).
(a) $P(A) = P(A \cap U) + P(A \cap U^c) = .20 + .18 = .38$.
(b) $P(U|A) = \dfrac{P(U \cap A)}{P(A)} = \dfrac{.20}{.38} = \dfrac{10}{19}$.

Remark.
Sometimes it is easier to compute an unconditional probability by computing it in pieces by assuming an event occurs. This is the situation where $P(A)$ might be difficult but, for some well-chosen B_i's, $P(A|B_i)$ is fairly easy/straightforward. This idea works especially well in situations where the experiment is performed in stages (i.e., a "sequential experiment").

The law of total probability (LOTP).
If $B_1, B_2, \ldots, B_n$ are mutually exclusive and exhaustive, then for any event $A \subset \Omega$,

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A \cap B_i) \\ &= \sum_{i=1}^{n} P(A|B_i)P(B_i) \end{aligned}$$

Advice:
The law of total probability is one of the more widely used results in all of applied probability and is an indispensable tool that will be used throughout this course. Read the examples I do, and know this result!

Here's a picture of this result with a partition of $n = 6$ events $B_1, B_2, \ldots, B_6$:

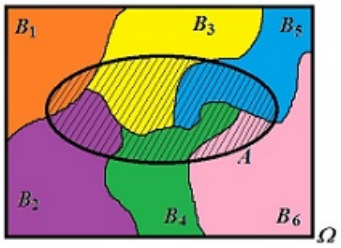


Figure. $B_1, \ldots, B_6$ partition Ω

The picture above shows the event A being decomposed into 6 "pieces": the piece where A meets B_1, the piece where A meets B_2, and so on. Since B_i for $i = 1, 2, 3, 4, 5, 6$ are mutually exclusive, $A \cap B_i$ for $i = 1, 2, 3, 4, 5, 6$ are also mutually exclusive; therefore, since

$$A = \bigcup_{i=1}^{n} (A \cap B_i),$$

by the countable additivity axiom $P(A) = \sum_{i=1}^{n} P(A \cap B_i)$. Moreover, since $P(A \cap B_i) = P(A|B_i)P(B_i)$, we also have the alternate form of the law of total probability.

Example.
We have two coins in a box. One coin is fair, the other has two heads. One of the coins is selected uniformly at random and tossed 3 times. Compute the probability you will get 3 heads.

SOLUTION:
Let F be the event we select the fair coin, F^c the biased coin, and let A be the event we get 3 heads. We are told $P(F) = \frac{1}{2}$ and $P(F^c) = \frac{1}{2}$, $P(A|F) = \frac{1}{8}$ and $P(A|F^c) = 1$. So,

$$P(A) = P(A|F)P(F) + P(A|F^c)P(F^c) = \frac{1}{8} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{9}{16}.$$

Example.
You toss a fair coin. If it shows heads, you roll *one* fair 6-sided die. If it shows tails, you roll *two* fair 6-sided dice. Compute the probability the sum total of the die(dice) is 6.

SOLUTION:
Let H be the event you toss heads, and let S be the event the total on the die(dice) shows 6. $P(S|H) = \frac{1}{6}$, and $P(S|H^c) = \frac{5}{36}$. Therefore,

$$P(S) = P(S|H)P(H) + P(S|H^c)P(H^c) = \frac{1}{6} \cdot \frac{1}{2} + \frac{5}{36} \cdot \frac{1}{2} = \frac{11}{72}.$$

Example.
You roll a fair 6-sided die three times. Find the probability the sum is 9.

SOLUTION:
There are other ways to do this problem, but I will illustrate a solution using the law of total probability, where I condition on the value of the first die. When we roll the first die, any of the numbers $1, 2, 3, 4, 5$ or 6 can occur equally likely. Let F_i be the event the first die shows an i. Let A be the event the sum of the three dice is 9. We want to compute $P(A)$.

We try to compute $P(A|F_i)$ for each $i = 1, 2, 3, 4, 5$ and 6. Start with $P(A|F_1)$. If the first die shows a 1, then the other two dice would need to show a total of 8 in order for the sum of the three dice to be 9; therefore $P(A|F_1) = P(8 \text{ on two dice}) = \frac{5}{36}$. Similarly, for $P(A|F_2)$, if a 2 occurs on the first die, then the other two dice would need to total to 7, so $P(A|F_2) = P(7 \text{ on two dice}) = \frac{6}{36}$. Continuing in this manner, $P(A|F_3) = P(6 \text{ on two dice}) = \frac{5}{36}$, $P(A|F_4) = P(5 \text{ on two dice}) = \frac{4}{36}$, $P(A|F_5) = P(4 \text{ on two dice}) = \frac{3}{36}$, and $P(A|F_6) = P(3 \text{ on two dice}) = \frac{2}{36}$. Since $P(F_i) = \frac{1}{6}$ for all i,

$$P(A) = \sum_{i=1}^{6} P(A|F_i)P(F_i) = \frac{1}{6}\left(\frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36}\right) = \frac{25}{216}.$$

Example. (Simple epidemic model)
A population of size $N = 10$ consists of people who are either diseased or non-diseased. Within the population a subset of size $D = 3$ are diseased ($N - D = 7$ are non-diseased). We randomly select a subset of size 2. If they are both non-diseased, put them back into the population. If you get one of each, the non-diseased person becomes diseased and then *both are put back into the population*. If they are both diseased, remove them from the population. *One* person is then selected uniformly at random from the group. Compute the probability they are diseased.

SOLUTION:
You recognize this experiment is performed in two stages. In the first stage we select a subset of size 2, and the configuration of the population changes depending on the sample that was chosen.

Let D_i denote the event that, in the first stage, we select i diseased people. If $i = 0$, then the population configuration remains the same (two just put the two non-diseased people back into the population). If $i = 1$, then the population remains the same size, namely, N, but the number of diseased people increases from D to $D + 1$ (non-diseased goes from $N - D$ to $N - D - 1$). If $i = 2$, then the population decreases from N to $N - 2$, but the number of diseased people also decreases from D to $D - 2$.

In the second stage we select one person from the resulting population. Let I be the event that we select a diseased person.

$$P(I|D_0) = \tfrac{D}{N} = \tfrac{3}{10}, \qquad P(I|D_1) = \tfrac{D+1}{N} = \tfrac{4}{10}, \qquad P(I|D_2) = \tfrac{D-2}{N-2} = \tfrac{1}{8}.$$

$$P(D_0) = \frac{\binom{D}{0}\binom{N-D}{2}}{\binom{N}{2}} = \tfrac{7}{15}, \quad P(D_1) = \frac{\binom{D}{1}\binom{N-D}{1}}{\binom{N}{2}} = \tfrac{7}{15}, \quad P(D_2) = \frac{\binom{D}{2}}{\binom{N}{2}} = \tfrac{1}{15}.$$

By the law of total probability

$$P(I) = \frac{3}{10} \cdot \frac{7}{15} + \frac{4}{10} \cdot \frac{7}{15} + \frac{1}{8} \cdot \frac{1}{5} = \frac{211}{600} \approx 0.35.$$

Exercise for the student.
Try to re-do this example with one small change: in the case $i = 1$ where we draw 1 diseased and 1 nondiseased, instead of making the non-diseased person diseased, flip a fair coin so that with probability $p = \frac{1}{10}$ make the non-diseased person diseased; otherwise, keep them non-diseased. Then put them both back into the population.

Example.
In a standard deck, a "10-card" is card with rank 10, Jack, Queen or King. I deal you two cards from a standard deck. If you make '21' (i.e., an Ace together with a 10-card) you win, and the games stops. Else, you get to trade in either 1 or 2 cards to make 21. You trade one card if you need exactly one of an Ace or a 10-card to get 21. You trade 2 cards if you have neither an Ace nor a 10-card. What's the chance you win?

SOLUTION:
Let W be the event you win, i.e., make 21. Let T_0 be the event you trade 0, meaning you made 21 on the initial deal, T_A the event that you trade in one card needing an Ace, T_{10} the event that you trade in one card needing a 10-card, and T_2 the event that you trade in 2 cards. In this problem I assume if you haven't made 21 on the initial deal you will *not* trade in an Ace nor will you trade in the 10-card.

$P(W|T_0) = 1$.

$P(W|T_A) = \frac{\binom{4}{1}}{\binom{50}{1}}$ since there are 4 Aces in the remaining deck of 50 cards.

$P(W|T_{10}) = \frac{\binom{16}{1}}{\binom{50}{1}}$ since there are 16 10-cards in the remaining deck of 50 cards.

$P(W|T_2) = \frac{\binom{4}{1}\binom{16}{1}}{\binom{50}{2}}$.

$P(T_0) = \frac{\binom{4}{1}\binom{16}{1}}{\binom{52}{2}}$ since there are 4 Aces and 16 10-cards.

$P(T_A) = \frac{\binom{16}{2}+\binom{16}{1}\binom{32}{1}}{\binom{52}{2}}$: you have either two of 16 10-cards or one of the 10-cards and one of the remaining 32 non-10-cards and non-Aces.

$P(T_{10}) = \frac{\binom{4}{2}+\binom{4}{1}\binom{32}{1}}{\binom{52}{2}}$: you have either 2 Aces or one of the 4 Aces and one of the remaining 32 non-10-cards and non-Aces.

$P(T_2) = \frac{\binom{32}{2}}{\binom{52}{2}}$: you drew 2 cards from the 32 non-Aces and non-10-cards.

By the law of total probability,

$$\begin{aligned} P(W) &= P(W|T_0)P(T_0) + P(W|T_A)P(T_A) + P(W|T_{10})P(T_{10}) + P(W|T_2)P(T_2) \\ &= 1 \cdot \frac{\binom{4}{1}\binom{16}{1}}{\binom{52}{2}} + \frac{\binom{4}{1}}{\binom{50}{1}} \cdot \frac{\binom{16}{2} + \binom{16}{1}\binom{32}{1}}{\binom{52}{2}} + \frac{\binom{16}{1}}{\binom{50}{1}} \cdot \frac{\binom{4}{2} + \binom{4}{1}\binom{32}{1}}{\binom{52}{2}} + \frac{\binom{4}{1}\binom{16}{1}}{\binom{50}{2}} \cdot \frac{\binom{32}{2}}{\binom{52}{2}} \\ &\approx 0.138 \end{aligned}$$

Example.
You roll a fair 6-sided die repeatedly. Compute the probability that no odd number occurs before the first 6 appears.

SOLUTION:
Where do we start? Think about what sample points must look like in this event. Look at where the first 6 can possibly occur, and then we need to make sure no odds occur before it (evens other than 6 are okay, but $1, 3$ and 5 are no-no's).

$$6, \qquad \underbrace{26, \quad 46}_{\text{1st 6 on 2nd roll}} \quad , \quad \underbrace{226, \quad 246, \quad 426, \quad 446,}_{\text{1st 6 on 3rd roll}} \quad \text{etc.}$$

If we knew where the first 6 occurs, say on the i roll ($i \geq 1$), then on rolls 1 thru $i-1$ we can only have 2's and 4's repetition allowed. This suggests that we condition on which roll the first 6 occurs. Let F_i be the event that the first 6 occurs on the ith roll. Let A be the event that only 2's and 4's occur before the first 6 (i.e., no odds!).

By the law of total probability

$$\begin{aligned} P(A) &= P(A \cap F_1) + P(A \cap F_2) + P(A \cap F_3) + P(A \cap F_4) + \cdots \\ &= \frac{1}{6} + \frac{2}{6^2} + \frac{2^2}{6^3} + \frac{2^3}{6^4} + \cdots \qquad \text{(a geometric series with } r = \tfrac{1}{3}\text{)} \\ &= \frac{\frac{1}{6}}{1 - \frac{1}{3}} = \frac{1}{4}. \end{aligned}$$

Remark. (some high-powered intuition)
There is an interesting intuitive argument for why the answer to this problem is $\frac{1}{4}$. It relies on the fact that when you roll a die repeatedly *every* one of the six numbers $1, 2, 3, 4, 5$ and 6 will eventually occur, i.e., we can ignore extreme instances where a sequence is missing at least one of the numbers. The intuitive idea is then: when we look at where the first 6 occurs in a sequence relative to the first 1, the first 3 and the first 5; then each of the 4! permutations of the numbers $1, 3, 5$, and 6 should be equally likely. That is, when we look at an infinite sequence of rolls, and we identify where the first 1, the first 3, the first 5 and the first 6 occurs within it, then each of the 4! possibilities should be equally likely. Therefore, of the 4! equally likely ways of seeing these numbers appear for their first time, 3! of them keep the 6 before the $1, 3$ and 5, so the answer is $\frac{3!}{4!} = \frac{1}{4}$.

The Bayes rule.

This section is just a straightforward extension of the law of total probability. The situation now is like this: We have an event A and a partition $B_1, B_2, \ldots, B_n$ of Ω. We know $P(A|B_i)$ and $P(B_i)$ for $i = 1, 2, \ldots, n$ and we want $P(B_j|A)$ for some fixed j. But, this is not hard to find. Look:

$$\begin{aligned} P(B_j|A) &= \frac{P(B_j \cap A)}{P(A)} \qquad \text{(by the conditional probability formula)} \\ &= \frac{P(A|B_j)P(B_j)}{P(A)} \qquad \text{(by the multiplicative rule on page 77)} \\ &= \frac{P(A|B_j)P(B_j)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots P(A|B_n)P(B_n)}. \qquad \text{(by LOTP)} \end{aligned}$$

We just proved the formula called...

Bayes rule.
If $B_1, B_2, \ldots, B_n$ are mutually exclusive and exhaustive, and $A \subseteq \Omega$ is any event, then for any j,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}.$$

Advice:
Don't memorize the Bayes formula: it's usually better to understand the computation of $P(B_j|A)$ above.

We now reconsider some old examples we did from the law of total probability section.

Example.
We have two coins in a box. One coin is fair, the other has two heads. One of the coins is selected uniformly at random and tossed 3 times. *Old question*: Compute the probability you will get 3 heads.
New question: Given you tossed 3 heads, what's the probability you picked the fair coin?

SOLUTION:
$P(F|A) = \dfrac{P(F \cap A)}{P(A)}$, so we just need to compute $P(A)$ and $P(F \cap A)$. We can get $P(A)$ via the law of total probability:

$$P(A) = P(A|F)P(F) + P(A|F^c)P(F^c) = \frac{1}{8} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{9}{16}.$$

Moreover, by the multiplicative rule, $P(F \cap A) = P(A \cap F) = P(A|F)P(F) = \dfrac{1}{8} \cdot \dfrac{1}{2} = \dfrac{1}{16}$.
Therefore,

$$P(F|A) = \frac{P(A|F)P(F)}{P(A)} = \frac{\frac{1}{16}}{\frac{9}{16}} = \frac{1}{9}.$$

Example.
You toss a fair coin. If it shows heads, you roll *one* fair 6-sided die. If it shows tails, you roll *two* fair 6-sided dice. *Old question*: Compute the probability the sum total of the die(dice) is 6.
New question: If the sum total of the die(dice) is 6, what's the chance you flipped heads?

SOLUTION:
$P(H|S) = \dfrac{P(H \cap S)}{P(S)} = \dfrac{P(S|H)P(H)}{P(S)}$, where, by the law of total probability,

$$P(S) = P(S|H)P(H) + P(S|H^c)P(H^c) = \frac{1}{6} \cdot \frac{1}{2} + \frac{5}{36} \cdot \frac{1}{2} = \frac{11}{72}.$$

Moreover, $P(S|H)P(H) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$; so, $P(H|S) = \frac{\frac{1}{12}}{\frac{11}{72}} = \frac{6}{11}$.

Example.
You roll a fair 6-sided die three times. *Old question*: Find the probability the sum is 9.
New question: If the sum is 9, what's the probability the first die shows a 1?

SOLUTION:
By law of total probability,

$$P(A) = P(A|F_1)P(F_1) + \cdots + P(A|F_6)P(F_6) = \frac{25}{216}.$$

Therefore,

$$P(F_1|S) = \frac{P(A|F_1)P(F_1)}{P(A)} = \frac{\frac{5}{36} \cdot \frac{1}{6}}{\frac{25}{216}} = \frac{1}{5}.$$

Example.
You roll a fair 6-sided die repeatedly. *Old question*: Compute the probability that no odd number occurs before the first 6 appears.
New question: If no odds occur before the first 6, what's the probability the first 6 occurs on the first roll?

SOLUTION:
We computed earlier that $P(A) = \frac{1}{4}$ and $P(F_1 \cap A) = P(F_1) = \frac{1}{6}$ since $F_1 \subseteq A$. Therefore,

$$P(F_1|A) = \frac{\frac{1}{6}}{\frac{1}{4}} = \frac{2}{3}.$$

Remark.
In all these examples, the new question asked for a conditional probability with the roles of the events involved being *reversed*.

Example.
At a certain factory 3 machines make widgets. Respectively, 30%, 50%, and 20% of the widgets are made by machines A, B, and C. Historically, it is known that 4%, 5%, and 3% of the widgets made by A, B, and C, respectively are defective. A widget is randomly sampled from all widgets made and it is found to be defective. Compute the probability it was made by machine A, then by machine B, and finally, by machine C.

SOLUTION:
Let A (resp., B, C) be the event the widget was made by machine A (resp., B,C), and D the event the widget is defective. We are told the following information:

$$P(A) = .3, \quad P(B) = .5, \quad P(C) = .2,$$

$$P(D|A) = .04, \quad P(D|B) = .05, \quad \text{and} \quad P(D|C) = .03.$$

We want to compute $P(A|D), P(B|D)$, and $P(C|D)$. I'll point out the obvious and state that a widget cannot be made by more than one machine, so A, B, and C are mutually exclusive; moreover, every widget is made by one of A, B, or C so they are exhaustive. Therefore, the law of total probability gives

$$\begin{aligned} P(D) &= P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C) \\ &= .04(.3) + .05(.5) + .03(.2) \\ &= .012 + .025 + .006 = .043, \end{aligned}$$

and, by Bayes rule,

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{.012}{.043} \approx 0.279$$
$$P(B|D) = \frac{P(D|B)P(B)}{P(D)} = \frac{.025}{.043} \approx 0.581$$
$$P(A|D) = \frac{P(D|C)P(C)}{P(D)} = \frac{.006}{.043} \approx 0.140.$$

In the Bayes rule setting, the conditional probabilities $P(D|A), P(D|B)$, and $P(D|C)$ are called the ***likelihoods***, the values $P(A), P(B)$, and $P(C)$ are called the ***prior probabilities*** and $P(A|D), P(B|D)$, and $P(C|D)$ are called the ***posterior probabilities***. Bayes rule give us a way to compute the posterior probabilities from the prior probabilities and the likelihoods.

Independence.

The concept of independence is unique to probability theory. We first discuss what it means for two events to be independent, then we discuss what it means for several (or infinite sequences of) events to be independent. Later in the course (after we've discuss random variables) we will revisit the independence idea and discuss what it means for random variables to be independent.

Independence of two events.
Suppose A and B are events. We say A and B are ***independent*** provided

$$P(A \cap B) = P(A)P(B).$$

If A and B are not independent, we say they are ***dependent***.

Remark. (intuitive meaning of independence of events)
Loosely speaking, A and B are independent means $P(A|B) = P(A)$ and $P(B|A) = P(B)$, i.e., knowledge of one of these events occurring does not influence the probability of the other event. To see why, assume $P(B) > 0$. Then, if A and B are independent we would have

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

Likewise, if $P(A) > 0$ and A and B are independent we would have $P(B|A) = P(B)$.

The reason we say "loosely speaking" is that we also want the definition of independence to hold for events that have probability *zero*, too, and strictly speaking, we can't use the conditional probability formula if the conditioning event has probability 0. Nevertheless, if the events have positive probability and we can show one of $P(A|B) = P(A)$ or $P(B|A) = P(B)$ then the events A and B are independent and if, in addition, $P(A) > 0$ and $P(B) > 0$ then the other will be true as well! Let's show this:

Suppose $P(A) > 0$ and $P(B) > 0$ and, without loss of generality, say $P(A|B) = P(A)$. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A) \implies P(A \cap B) = P(A)P(B) \implies A, B \text{ independent.}$$

Moreover, because A and B are independent and $P(A) > 0$ as well, it follows

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

and the other conditional statement holds, too!

Remark.
Suppose A and B are independent events. It seems intuitively clear that if knowledge that B occurs does not influence the probability of A, then knowledge that B doesn't occur shouldn't influence the probability of A either. In fact, more is true...

If A and B are independent events, then
A and B^c are independent events,
A^c and B are independent events, and
A^c and B^c are independent events.

In fact, let's prove: A and B independent implies A and B^c are independent. We need to show that $P(A \cap B^c) = P(A)P(B^c)$. To this end

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \quad \text{(used } A \text{ and } B \text{ being independent)} \\ &= P(A)[1 - P(B)] = P(A)P(B^c) \quad \text{(complementary rule).} \end{aligned}$$

Exercise for the student.
Please provide similar proofs to show that A and B independent implies the other pairs of events are independent, too.

Example.
Consider the following three Venn diagrams. Each shows probabilities within the designed regions. Which diagrams show independent events, which show dependent events? Explain.

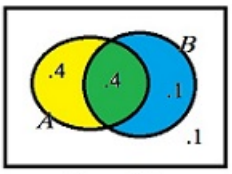


Figure (a)

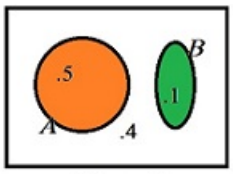


Figure (b)

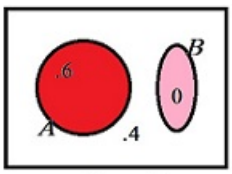


Figure (c)

SOLUTION:
Figure (a) shows independent events: $P(A) = .8$, $P(B) = .5$, and $P(A \cap B) = .4 = P(A)P(B)$. Notice that from the conditional probability point of view, if, for instance, given B occurred, then $A \cap B$ and $A^c \cap B$ occur in the same proportions as A and A^c in the original sample space: $\frac{.4}{.5} = .8$, $\frac{.1}{.5} = .2$. The information that B occurred didn't influence the probability of A (nor A^c).

Figure (b) shows dependent events: $P(A) = .5$, $P(B) = .1$, and $P(A \cap B) = P(\varnothing) = 0 \neq P(A)P(B)$. This is a situation where mutually exclusive events are dependent.

Figure (c) shows independent events: $P(A) = .6$, $P(B) = 0$, and $P(A \cap B) = P(\varnothing) = 0 = P(A)P(B)$. This is a special case where mutually exclusive events can be independent (when one or both have probability 0).

Example and discussion.
We have a box filled with 3 red and 4 green balls. We choose two balls. Let A be the event that the first ball is red. Let B be the event the second ball is green. Discuss whether or not these events are independent in each of the following situations:
sampling with replacement, and sampling without replacement.

SOLUTION:
First, try to guess the answer without any computations. In each situation we ask: Will information of one of these events change the probability of the other? Let's look at each situation separately...

If we sample with replacement, then the configuration of the box is the same at the second draw as it was for the first, so I'd expect these to be independent. However, if we were sampling without replacement, then knowledge of the first draw will now *alter* the configuration of box at the second draw, so we cannot expect these events to be independent.

In the situation of sampling *with replacement*:
$P(A) = \frac{3\cdot 7}{7\cdot 7} = \frac{3}{7}$ as there are 7^2 possible outcomes but only $3 \cdot 7$ of these ordered 2-tuples have a red in the first position. $P(B) = \frac{7\cdot 4}{7\cdot 7} = \frac{4}{7}$ as there are $7 \cdot 4$ ordered 2-tuples that have a green ball in the second position. and $P(A \cdot B) = \frac{3\cdot 4}{7\cdot 7} = \frac{3}{7} \cdot \frac{4}{7} = P(A)P(B)$ shows A and B are independent.

In the situation of sampling *without replacement*:
$P(B|A) = \frac{4}{6}$ since a red drawn on the first leaves us with 6 equally likely balls of which 4 are green, but then $P(B|A) \neq \frac{4}{7} = P(B)$ and, therefore, the events A and B must be dependent.

Exercise for the student.
Roll two fair 6-sided dice. Let A be the event that the sum total of the dice is even. Let B be the event the sum total is $3, 6, 9$ or 12. Are A and B independent?

Exercise for the student.
Flip a coin twice: $\Omega = \{hh, ht, th, tt\}$.
Let H_1 be the event of a head on the first toss.
Let H_2 be the event of a head on the second toss.
Verify whether or not H_1 and H_2 are independent in each of the following situations:

(a) Ω is equally likely, i.e., $P(\{hh\}) = P(\{ht\}) = P(\{th\}) = P(\{tt\}) = \frac{1}{4}$.
(b) $P(\{hh\}) = .06, P(\{ht\}) = .24, P(\{th\}) = .14, P(\{tt\}) = .56$.
(c) $P(\{hh\}) = .4, P(\{ht\}) = .3, P(\{th\}) = .2, P(\{tt\}) = .1$.

Before we state the definition of independence of *many* events, we first define what it means for just 3 events to be independent.

Independence of 3 events.
Let A, B, and C be events. We say they are ***(mutually) independent*** provided ***all*** 4 of the following conditions hold true:

$$\left.\begin{array}{rcl} P(A \cap B) & = & P(A)P(B) \\ P(A \cap C) & = & P(A)P(C) \\ P(B \cap C) & = & P(B)P(C) \end{array}\right\} \quad (*)$$

and

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

If A, B and C satisfy the conditions **(*)**, they are called ***pairwise independent***. If, additionally, they also satisfy the last condition, we call them *mutually* independent or, just, independent.

Remark.
From the definition above, 3 independent events are always pairwise independent, but the converse is not necessarily true. There are examples of events that are pairwise independent but not (mutually) independent. Here's a classic example:

Example. (pairwise independent events that are not independent)
Roll two fair 6-sided dice (one red, one green). Let
A be the event the red die shows a 1,
B be the event the green die shows a 6,
C be the event the sum is 7.
Show that A, B and C are pairwise independent but not mutually independent.

SOLUTION:
In this experiment we clearly have $P(A) = P(B) = P(C) = \frac{1}{6}$. Now,

$$A \cap B = \{16\}, \quad A \cap C = \{16\}, \quad \text{and} \quad B \cap C = \{16\}.$$

Therefore, $P(A \cap B) = \frac{1}{36} = P(A)P(B)$, and, similarly, $P(A \cap C) = P(A)P(C)$, $P(B \cap C) = P(B)P(C)$ and these events are pairwise independent. *However*, $A \cap B \cap C = \{16\}$ and

$$P(A \cap B \cap C) = \frac{1}{36} \neq P(A)P(B)P(C) = \frac{1}{216},$$

so these events are *not* mutually independent!

Independence of events.
We say a collection of events (finite or infinite) $\{A_1, A_2, A_3, \dots\}$ is ***independent*** provided *every* finite subcollection (of two or more) of these events,

$$\{A_{i_1}, A_{i_2}, \dots, A_{i_k}\},$$

has the property

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

Remark.
Generalizing an earlier result, if $\{A_1, A_2, A_3, \dots\}$ is a collection of independent events, then replacing any number of the A_i in this collection by its complement, A_i^c, will give another independent collection.

Remark.
In practice, independence can usually be assumed by virtue of the experiment. Here are some illustrations where assuming independence is plausible:

• Alan and Betty are each tossing coins. Then the result of Alan's tosses shouldn't influence Betty's tosses (and vice-versa).

• A person rolls a die repeatedly (or tosses a coin repeatedly). The the result on a roll (toss) shouldn't influence other rolls (tosses).

• A person buys a lottery ticket each day. The events she has a winning ticket on day i ($i = 1, 2, 3, \dots$) should be independent.

Remark.
The thing that is nice about independence (when we have it or when it can be assumed) is that, if $A_1, A_2, A_3, \dots$ are indepedent, then, for instance,

$$\begin{aligned}
P\left(\bigcup_{i=1}^{n} A_i\right) &= 1 - P\left(\left(\bigcup_{i=1}^{n} A_i\right)^c\right) && \text{(by complementary rule)} \\
&= 1 - P\left(\bigcap_{i=1}^{n} A_i^c\right) && \text{(by DeMorgan's law)} \\
&= 1 - \prod_{i=1}^{n} P\left(A_i^c\right) && \text{(by independence)} \\
&= 1 - \prod_{i=1}^{n} \left(1 - P(A_i)\right) && \text{(by complementary rule),}
\end{aligned}$$

which may be easier to employ than, say, the inclusion-exclusion rule.

Example. (The State of Maryland Powerball lottery)
A Powerball lottery ticket is formed in two steps.

step 1: select of 5 numbers from 1 thru 69 inclusive;

step 2: choose a Powerball number from 1 thru 26 inclusive.

The pair

$$\begin{pmatrix} \text{subset of size 5} & \text{choose one from} \\ \text{from } \{1,2,\ldots,69\}, & \{1,2,\ldots,26\} \end{pmatrix}$$

is a Powerball lottery ticket. You can play this game each Wednesday and Saturday every week until you die. As of now, the cost is \$2 per lottery ticket.
(a) (Old question) How many Powerball lottery tickets are possible?
(b) What's the probability you hold a winning ticket*?
(c) If you play twice per week for the next 50 years, what the probability you win at least once?
(d) Suppose that you and 4999 of your most trusted friends all *independently* follow this same strategy and agree to share the winnings if anyone wins. What's the probability you win at least once now?
* By *win* and *winning ticket* I mean winning the whole shebang – the jackpot!

SOLUTION:
(a) The basic counting principle implies there are $\binom{69}{5}26 = 292,201,338$ possible Powerball tickets.

(b) The reciprocal of the answer in part (a) is the probability of selecting the winning ticket: $\frac{1}{292201338} \doteq 3.422 \times 10^{-9}$.

(c) Playing twice per week for 50 years means you are buying

$$2\,\frac{\text{tickets}}{\text{week}} \times 52\,\frac{\text{weeks}}{\text{year}} \times 50\,\text{years} = 5200 \ \text{ lottery tickets.}$$

The event that we win at least once is

$$\bigcup_{i=1}^{5200} W_i,$$

where W_i is the event we win on the ith lottery ticket. By the way, we know that, for every i, $P(W_i) = P(W_1) \doteq 3.422 \times 10^{-9}$.
Therefore,

$$P\left(\bigcup_{i=1}^{5200} W_i\right) = 1 - \prod_{i=1}^{5200} (1 - P(W_i)) = 1 - \left(1 - 3.422 \times 10^{-9}\right)^{5200} \doteq 1.779 \times 10^{-5} \approx .0000179.$$

(d) Let B_1 be the event you win at least once with this strategy. Let B_i be the event that friend i wins at least once with this strategy for $i = 2, 3, 4, \ldots, 5000$. Then, assuming all $B_1, B_2, \ldots, B_{5000}$ are independent, the probability we win is

$$P\left(\bigcup_{i=1}^{5000} B_i\right) = 1 - \prod_{i=1}^{5000} (1 - P(B_i)) = 1 - \left(1 - 1.779 \times 10^{-5}\right)^{5000} \approx .085.$$

III. Random variables, discrete random variables.

Random variables

Up to know we've been dealing with probability at the experiment level, i.e., we've concerned ourselves with the sample space - either counting sample points or observing some property of the sample space and/or events.

In many cases we do not have to concern ourselves with the sample space, i.e., the sample points themselves. For example, many times an experiment is performed the experimenter might make a measurement on the outcome and observe, instead, some real number. In this way the experimenter may not care necessarily about the specific sample point observed but rather the set of sample points having a specific measured value (or values).

A ***random variable*** (or ***rv***) X is a real-valued function defined on all of Ω, $X : \Omega \to \mathbb{R}$.

Remark.
Chance picks the ω, and we observe $X(\omega)$ which is a real number. Think that X is making a "measurement" on the ω that chance produced. X is really a deterministic function - we call it a "random" variable because we don't typically observe the random input ω into the function and, therefore, it would appear the X randomly outputs $X(\omega)$.

Notation.
If $x \in \mathbb{R}$, then $(X = x)$ is shorthand for the subset of Ω that X maps to x:

$$(X = x) = \{\omega \in \Omega : X(\omega) = x\}.$$

This is called the ***preimage of*** $\{x\}$ under X. Mathematicians sometimes call this the ***inverse image of*** $\{x\}$ ***under*** X and denote it by $X^{-1}[\{x\}]$.

The following example demonstrates that rvs generalize experiment-level probability.

Example. (A Bernoulli rv)
Let $A \subseteq \Omega$ be any event. Then, by defining the rv

$$X(\omega) = \begin{cases} 1 & \text{for } \omega \in A \\ 0 & \text{for } \omega \in A^c \end{cases},$$

we would have $P(X = 1) = P(A)$ and $P(X = 0) = P(A^c)$.

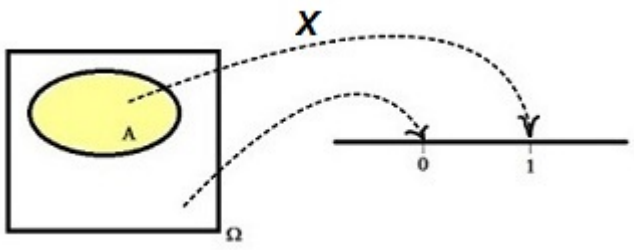


Figure. The random variable X is mapping Ω into the set $\{0, 1\}$.

The figure above shows that X is mapping all the $\omega \in A$ to the point 1 and mapping all the $\omega \in A^c$ to the point 0. In particular, if X returns the value 1, then we know that the $\omega \in A$, but we may not necessarily know which specific $\omega \in A$ occurred (nor do we care).

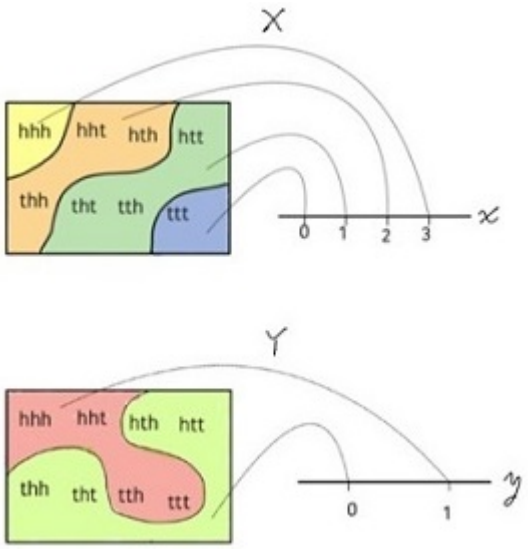
Advice:
When considering random variables (however they are defined) it's *always* a good idea to keep in mind the possible values the random variable can take on (and which they cannot).

Example.
Toss a fair coin 3 times: $\Omega = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$.
Consider the rvs:

X = the number of heads.

$Y = 1$ if the first two tosses are the same parity, and $Y = 0$ otherwise.

X takes possible values 0, 1, 2 and 3. Y takes possible values 0 and 1.
The set of possible values $\{0, 1, 2, 3\}$ is the ***image of*** X, $\{0, 1\}$ is the ***image of*** Y.
To the right of the figure below are the preimages.



$(X = 0) = \{ttt\}$
$(X = 1) = \{htt, tht, tth\}$
$(X = 2) = \{hht, hth, thh\}$
$(X = 3) = \{hhh\}$

$(Y = 0) = \{hth, htt, thh, tht\}$
$(Y = 1) = \{hhh, hht, tth, ttt\}$

Figure. Visualization of the X and Y mappings into the real numbers.

Remark.
Assume the coin is fair. We get the ***probability mass function (pmf)*** of rv X:

$$P(X = 0) = \tfrac{1}{8},\ P(X = 1) = \tfrac{3}{8},\ P(X = 2) = \tfrac{3}{8} \text{ and } P(X = 3) = \tfrac{1}{8}.$$

We can also write this in ***tabular form***:

x	0	1	2	3
$P(X = x)$	.125	.375	.375	.125

There happens to be a ***functional form*** as well:

$$P(X = x) = \frac{\binom{3}{x}}{8} \quad \text{for } x = 0, 1, 2, 3.$$

The pmf for Y is $P(Y = 0) = \frac{1}{2} = P(Y = 1)$.

Basic types of random variables (rvs).
We'll see that rvs are extremely useful in helping us model common experiments and quantities that come up in practice. We now talk about types of rvs we will encounter.

There are essentially two different types of rvs: discrete rvs and continuous rvs, and we classify the type of rv we are dealing with by investigating the set of possible values the rv can take on. A rv is classified as ***discrete*** if the set of possible values it can take on forms a discrete set, i.e., a finite or countably infinite set of (real) values. The criterion for classifying a rv as ***continuous*** is a little tougher; loosely speaking, the set of possible values is a "continuum", i.e., either an interval (a, b) (or $[a, b]$) of the real line with $a < b$ (allowing the possibility that $a = -\infty$ and/or $b = +\infty$ here) or a union of such intervals. But please read the next paragraph.

A rv can be neither discrete nor continuous; for example, a rv whose set of possible values is $\{x \in \mathbb{R} : 0 \leq x \leq 1\} \cup \{2, 3, 4\}$ - this is not a discrete set, it is also not a union of intervals described above. Another example of a rv that is neither discrete nor continuous is when the rv is a ***mixture*** of the two types of rvs. For example, say we roll a fair die. If we roll an even number, toss a fair coin one time, and set $X = 1$ if you get a head, $X = 0$ if you get a tail (a discrete rv); else, if we roll an odd number, pick a number uniformly at random from the interval $[0, 1]$ so that $0 \leq X \leq 1$ (a continuous rv). The rv X just described takes only values in the interval $[0, 1]$ but is not continuous. Neither is it discrete. The rv just described is a mixture of discrete and continuous rvs. Random variables of this type are best understood using their so-called ***cumulative distribution functions***. We will discuss this later in the course.

Lastly, I'll mention that continuous random variables can be further subcategorized into ***absolutely continuous*** and ***singular continuous***. The distinction between these types is of no importance right now, but when we get into a discussion later in the course on continuous random variables we'll be sure to mention the differences then.

Exercise.
Classify each of the following random variables, and determine their set of possible values.
(a) Roll two 6-sided dice. S is the sum of the upfaces.
(b) A dart lands on a dartboard*. R is the distance from the center where a dart lands.
(c) Throw two darts at a dartboard. If both darts land on the dartboard, then X is the distance between x-coordinate values where these darts land; else X returns the number of darts that miss the dartboard.

* Assume all dartboards are represented as the unit disk: $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$.

Discrete random variables.

A rv X is called ***discrete*** if the ***image of*** X:

$$\{x \in \mathbb{R} : X(\omega) = x \text{ for some } \omega \in \Omega\}$$

is a discrete set, i.e., finite or countably infinite. The image of X is the set of possible real values the random variable can return.

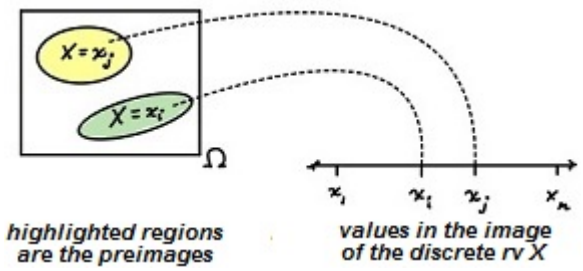


Figure. Images and preimages of the discrete rv X.

Some facts regarding random variables.
Fact 1: The events $(X = x)$ for $x \in \text{Image}(X)$ are mutually exclusive subsets of Ω: $(X = x) \cap (X = y) = \varnothing$ when $x \neq y$ (it is impossible for X to map the same ω to two different values).
Fact 2: The events $(X = x)$ for $x \in \text{Image}(X)$ are exhaustive: $\bigcup_{x \in \text{Image}(X)} (X = x) = \Omega$.

These facts imply the events $(X = x)$ for $x \in \text{Image}(X)$ are ***mutually exclusive and exhaustive***. The axioms imply

$$\sum_x P(X = x) = 1,$$

where the sum is over all possible values of the random variable X, i.e., its image. The function of x, $p_X(x) := P(X = x)$, is called the ***probability mass function (pmf) of*** X. The image of X is the ***support of this pmf***.

Computing probabilities using pmfs.
Once we know the pmf of a discrete rv, computing the probability of any event involving this one rv is straightforward: If $I \subseteq \mathbb{R}$, then

$$P(X \in I) = \sum_{x \in I} P(X = x),$$

i.e., sum the probability masses for each value in the support that belongs to I. In this way, we no longer need to compute such probabilities at the experiment level; ***the pmf allows us to essentially forget about the sample space when computing probabilities.***

Example.
Roll two fair 6-sided dice. Let X be the total (sum) showing on the upfaces. X is discrete since the image of X, namely,

$$\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

is a discrete set. For the record, the ***preimages*** for each value in the image are given in the following table:

$$\begin{array}{ll}
(X = 2) = \{(1,1)\} & (X = 3) = \{(1,2),(2,1)\} \\
(X = 4) = \{(1,3),(2,2),(3,1)\} & (X = 5) = \{(1,4),(2,3),(3,2),(4,1)\} \\
(X = 6) = \{(1,5),(2,4),(3,3),(4,2),(5,1)\} & (X = 7) = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \\
(X = 8) = \{(2,6),(3,5),(4,4),(5,3),(6,2)\} & (X = 9) = \{(3,6),(4,5),(5,4),(6,3)\} \\
(X = 10) = \{(4,6),(5,5),(6,4)\} & (X = 11) = \{(5,6),(6,5)\} \\
(X = 12) = \{(6,6)\} &
\end{array}$$

and, if $x \notin \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, then $(X = x) = \varnothing$.

Here's the pmf of X in different forms:

In ***tabular form***:

x	2	3	4	5	6	7	8	9	10	11	12
$p_X(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

In ***functional form***:

$$p_X(x) = P(X = x) = \frac{6 - |x - 7|}{36} \quad \text{for } x = 2, 3, 4, \ldots, 11, 12.$$

Compute each of the following:

$P(\underbrace{3 \leq X \leq 5}_{3,4,5}) = P(X = 3) + P(X = 4) + P(X = 5) = \frac{9}{36} = \frac{1}{4}$.

$P(\underbrace{2.5 \leq X \leq 5.5}_{3,4,5}) = \frac{1}{4}$.

$P(\underbrace{3 < X < 5}_{4}) = \frac{4}{36} = \frac{1}{9}$.

$P(\underbrace{5X - X^2 \geq 4}_{2,3,4}) = P(X = 2) + P(X = 3) + P(X = 4) = \frac{6}{36} = \frac{1}{6}$.

Random variables are extremely helpful in modeling common experiments that come up all the time in practice. In fact, some of these experiments are so common that the pmfs and their associated rvs are given special names.

Some named pmfs (discrete probability distributions) we will study:
The Bernoulli
The hypergeometric
The binomial
The Poisson
The geometric
The negative binomial
The discrete uniform
The multivariate hypergeometric
The multinomial

The term ***distribution*** is used here to mean the way the probability masses are distributed among its possible values.

1. The Bernoulli(p) distribution.
This is the probability (mass) distribution of a discrete random variable X that only takes two values 0 and 1, and

$$P(X = 1) = p \quad \text{and} \quad P(X = 0) = 1 - p.$$

Notation: We use the notation

$$X \sim \text{Bernoulli}(p)$$

to mean the random variable X has the Bernoulli(p) distribution shown above.

We can think about the Bernoulli(p) random variable this way:

We have a collection of objects of which a proportion p are labeled ***successes*** (the remaining proportion $1-p$ are labeled ***failures***). From this collection we randomly draw one object and set $X = 1$ if we draw a success, and set $X = 0$ if we draw a failure. This is the simplest random variable, and as such it is often used as a building block in modeling more complicated experiments... most commonly, experiments that are conducted in (possibly many) trials where on each trial we have a Bernoulli(p) random variable.

Here's an experiment we can model using Bernoulli(p) building blocks:

2. The hypergeometric distribution.
Suppose we have a ***finite*** collection of distinct objects, say N total, of which M are labeled successes (and, $N - M$ are failures). You can imagine a bag of $N = 100$ m&m candies of which $M = 20$ are colored red (and $N - M = 80$ are not red). We draw n objects from this collection one-at-a-time ***without replacing*** them after they are drawn. Obviously, $n \leq N$ else there is nothing left to draw.

In this experiment, we are interested in the random variable:
$X =$ the number of successes we draw in the n trials.

It should be clear that the set of possible values x of X must be between 0 and n inclusive, but also must satisfy

$$x \leq M \quad \textbf{\textit{and}} \quad n - x \leq N - M,$$

since the number of successes cannot exceed the total number of successes in the collection and the number of failures cannot exceed the total number of failures in the collection.

Therefore, for $x = 0, 1, 2, \ldots, n$, $x \leq M$ and $n - x \leq N - M$, the event $(X = x)$ doesn't depend on the order in which that n objects were selected, and since each selection of n objects is equally-likely, we have the ***pmf of the hypergeometric***:

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

Once we know the pmf, recall from page 97 computing probability is straightforward.

Example.
In a bag of 100 m&m's of which 20 are red, what's the probability that, in a randomly chosen handful of 5 candies, you have 1 or 2 red m&m's? At least one red m&m?

Let X count the number of red m&m's. Then

$$P(1 \leq X \leq 2) \;=\; P(X = 1) + P(X = 2) = \frac{\binom{20}{1}\binom{80}{4} + \binom{20}{2}\binom{80}{3}}{\binom{100}{5}} \approx 0.6275.$$

$$P(X \geq 1) \;=\; 1 - P(X = 0) = 1 - \frac{\binom{20}{0}\binom{80}{5}}{\binom{100}{5}} \approx 1 - 0.3193 = 0.6807.$$

In the hypergeometric experiment, i.e., sampling without replacement from a finite population of successes and failures, the underlying Bernoulli rvs were ***dependent*** because knowledge of what is drawn on a trial (success or failure) influences the probability on other trials. We now consider a similar experiment but where it happens that the Bernoulli sequence will be ***independent***.

3. The binomial(n, p) distribution.

Consider a sequence of n independent Bernoulli(p) trials, i.e., the result of each trial is either a 1 (success) or 0 (failure), and knowledge of which occurred will not influence the probabilities on other trials. We are interested in the rv

X = number of successes in n trials.

The ***pmf of binomial***(n, p) is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.$$

(Why?)

Represent the sample space of this experiment as the set of all n-tuples of 0's and 1's:

$$\Omega = \big\{(x_1, x_2, \ldots, x_n) : \text{ each } x_i \in \{0, 1\}\big\},$$

where a 1 in the ith entry means a success on trial i, a 0 in the ith entry means a failure in trial i. The event $(X = x) = \big\{\omega = (x_1, x_2, \ldots, x_n) : \text{ each } x_i \in \{0, 1\},\ \sum_{i=1}^n x_i = x\big\}$, i.e., it is the subset of n-tuples having x successes (1's) and $n - x$ failures (0's). Notice that

$$|(X = x)| = \binom{n}{x} = \frac{n!}{x!(n-x)!},$$

since $(X = x)$ is all the anagrams of the n-letter "word" comprised of x 1's and $n - x$ 0's. Moreover, each $\omega = (X = x)$ has the *same* probability $p^x(1-p)^{n-x}$ since the results on each trial are independent. It then follows that

$$P(X = x) = \sum_{\omega \in (X=x)} p^x (1-p)^{n-x} = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n.$$

Remark.
Suppose we have an infinite population of successes and failures with a proportion p of successes. A binomial(n, p) experiment can be thought of as sampling n objects from this population without replacement. Sampling without replacement in this population will lead to ***independent*** Bernoulli(p) trials. This is because knowledge of what was drawn will now ***not*** change the proportion of successes on later trials.

Another way to think about the binomial(n, p) experiment is having a ***finite*** collection of successes and failures but sample ***with replacement*** instead. Then the Bernoulli(p) trials will be independent here because knowledge of what type was drawn will not change the proportion of successes when we randomly draw on other trials (because we replace what was drawn).

Some interesting binomial experiments/rvs.

- Flip a fair coin n times. X counts number of heads. $X \sim \text{binom}(n, \frac{1}{2})$.
  This is the prototypical binomial experiment.
- Roll a fair 6-sided die 5 times. X is the number of times you roll a 1 or 2.
  $X \sim \text{binom}(5, \frac{1}{3})$.
- 90% of the eggs a hen lays are grade-A. The hen lays a batch of 20 eggs.
  X is the number of grade-A eggs in the batch. $X \sim \text{binom}(20, .9)$.
- You walk into a classroom with 50 other people, let X be the number of these people having your birthday. $X \sim \text{binom}(50, \frac{1}{365})$.
- Every Wednesday for a year you buy a scratch-off lottery ticket that claims there's a 5% chance you'll win a prize of \$50 or more on each ticket. Let X be the number of times you win \$50 or more in the year. $X \sim \text{binom}(52, .05)$.

Thought exercise.
In the above examples you should really try to understand either why the conditions of a binomial are plausible or what the assumptions of the binomial mean in the context of the example, what the identical "trials" are, what a "success" is, what the probability of success is on each trial.

For instance, in the hen example, the "trials" are each of the 20 eggs – they can be grade-A (success) or *not* grade-A (failure), the chance the hen lays a grade-A egg (i.e., probability of success on a trial) is 0.9, we're assuming that the quality of an egg in the batch does not influence the quality of any other egg in the batch, i.e., the quality from egg to egg are independent.

Examples.
Flip a fair coin 100 times. Compute the probability of exactly 50 heads. Of between 35 and 65 (inclusive) heads.

$P(X = 50) = \binom{100}{50}\left(\frac{1}{2}\right)^{100} \approx 0.08$.

$P(35 \le X \le 65) = \sum_{x=35}^{65}\binom{100}{x}\left(\frac{1}{2}\right)^{100} \approx 0.99821$ using Microsoft Excel©.

Roll a fair 6-sided die 5 times. Compute the probability a 1 or 2 occurs on exactly 3 dice.

$P(X = 3) = \binom{5}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2 = 10 \cdot \frac{1}{27} \cdot \frac{4}{9} = \frac{40}{243} \approx 0.1646$.

90% of the eggs a hen lays are grade-A. The hen lays a batch of 20 eggs. Compute the probability there is at least 19 grade-A eggs.

$P(X \ge 19) = \binom{20}{19}(.9)^{19}(.1)^1 + \binom{20}{20}(.9)^{20}(.1)^0 \approx 0.3917$.

You walk into a classroom with 50 other people. Compute the probability at least 1 person has your birthday.

$P(X \ge 1) = 1 - P(X = 0) = 1 - \binom{50}{0}\left(\frac{1}{365}\right)^0\left(\frac{364}{365}\right)^{50} = 1 - \left(\frac{364}{365}\right)^{50} \approx 0.128$.

You buy a scratch-off lottery ticket with a 5% win probability every Wednesday for a year. Compute the probability you lose on every ticket.

$P(X = 0) = \binom{52}{0}(.05)^0(.95)^{52} = .95^{52} \approx 0.069$.

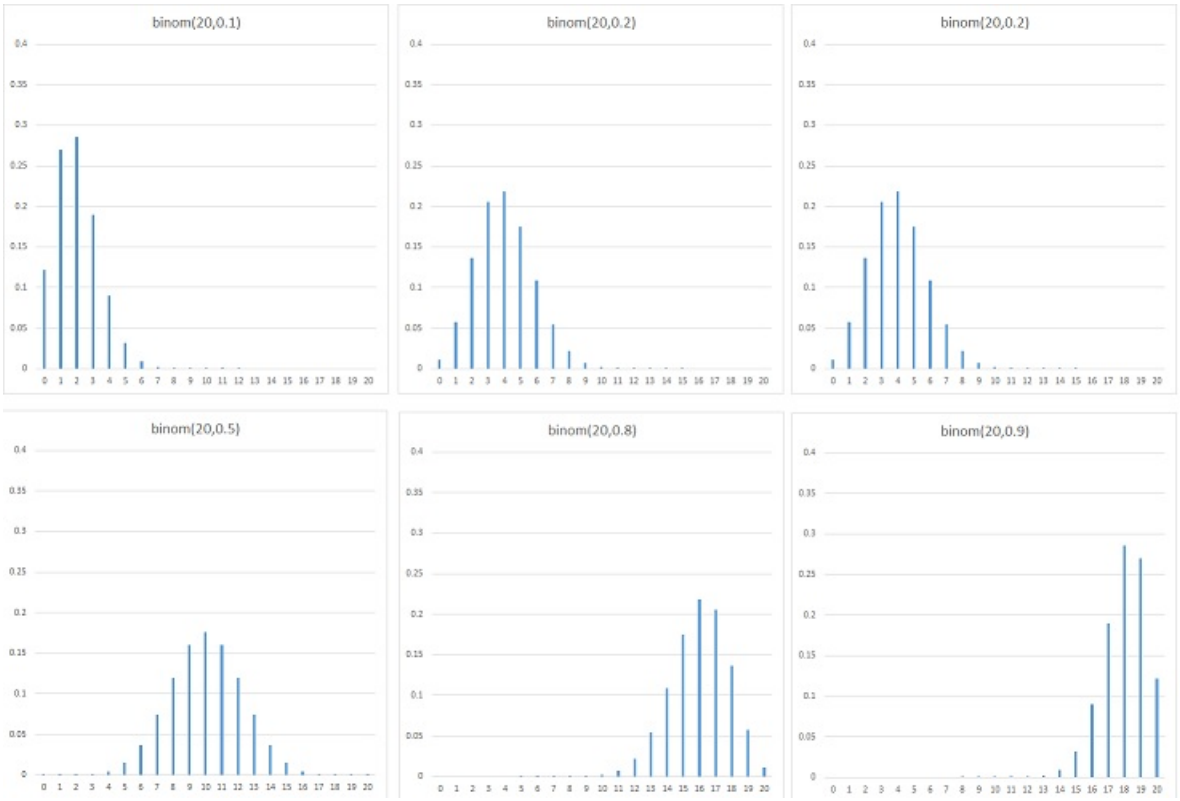


Figure. Some plots of binomial pmfs.

Calculus fact: the binomial theorem.
For integer $n \geq 1$ and real constants a and b,

$$\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a+b)^n.$$

Since $(a+b)^n = (b+a)^n$, we also have

$$\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k = (a+b)^n.$$

Advice:
The binomial theorem is absolutely indispensable - and will not only be used several times in this course but also the student will need to be able to recognize that they can use this result as the situation arises. ***Know this theorem and memorize it!***

Here's one use of the binomial theorem:

The binomial(n,p) pmf *is* a pmf:
Let $0 < p < 1$. By taking $a = p$ and $b = 1 - p$ in the binomial theorem we obtain

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

which shows the binom(n,p) probabilities sum to 1 as they should.

Calculus fact: general little "oh" notation.
We say a given function $f(x)$ of x ***is*** $\boldsymbol{o}(g(x))$ ***as*** $x \to c$ (pronounced "is little oh of $g(x)$ as x goes to c") if

$$\lim_{x \to c} \frac{f(x)}{g(x)} = 0.$$

The value c is allowed to be $+\infty$ or $-\infty$ as well. The little oh condition says that as x approaches c, the magnitude of the function $g(x)$ increases more rapidly than the magnitude of the function $f(x)$ in such a way that the ratio $f(x)/g(x)$ tends to 0. From the definition it should be clear that if a function $h(x)$ is $o(1)$ as $x \to c$, then $\lim_{x \to c} h(x) = 0$.

Exercise.
Please verify for yourself that
$\frac{1}{x^{3/2}} = o(1)$ as $x \to \infty$,
$\frac{1}{x^{3/2}} = o(\frac{1}{x})$ as $x \to \infty$,
$\frac{1}{x \ln(x)} = o(\frac{1}{x})$ as $x \to \infty$,
$x^5 = o(x^6)$ as $x \to \infty$, and
$x^2 - 1 = o(x + 1)$ as $x \to 1$.

I'll do the first two:
Certainly $\dfrac{\frac{1}{x^{3/2}}}{1} = \dfrac{1}{x^{3/2}} \to 0$ as $x \to \infty$. Therefore, $\frac{1}{x^{3/2}} = o(1)$ as $x \to \infty$.
Since $\dfrac{\frac{1}{x^{3/2}}}{\frac{1}{x}} = \dfrac{1}{x^{1/2}} \to 0$ as $x \to \infty$, $x^{-3/2} = o(1/x)$ as $x \to \infty$.
However, notice it is not $o(\frac{1}{x^2})$ as $x \to \infty$, nor is it $o(\frac{1}{x})$ as $x \to \mathbf{0}$.

Stirling's approximation:

$$m! = \sqrt{2\pi}\, m^{m+\frac{1}{2}} e^{-m}(1 + o(1)) \quad \text{as } m \to \infty,$$

or, equivalently,

$$\lim_{m\to\infty} \frac{m!}{\sqrt{2\pi}\, m^{m+\frac{1}{2}} e^{-m}} = \lim_{m\to\infty} (1 + o(1)) = 1.$$

This result says the relative error in approximating $m!$ by $\sqrt{2\pi}m^{m+\frac{1}{2}}e^m$ goes to 0 as m tends to ∞:

$$\frac{m! - \sqrt{2\pi}\, m^{m+\frac{1}{2}} e^{-m}}{\sqrt{2\pi}\, m^{m+\frac{1}{2}} e^{-m}} = o(1) \quad \text{as } m \to \infty.$$

Example.
Find the relative error in approximating 10! by Stirling's approximation.

Since $10! = 3628800$ and $\sqrt{2\pi}10^{10.5}e^{-10} = 3598695.619\ldots$, we find that

$$\frac{10! - \sqrt{2\pi}10^{10.5}e^{-10}}{\sqrt{2\pi}10^{10.5}e^{-10}} = 0.00836535\ldots.$$

Exercise.
Show that the Stirling's approximation to $\binom{N}{n}$ is

$$\frac{(N-n)^n \left(1 - \frac{n}{N}\right)^{-1/2}}{n!}(1 + o(1)).$$

Challenging exercise.
The remark on page 101 should lead us to believe that if a finite population is ***very*** large, then the hypergeometric and binomial distributions should be close. Fix integers x and n such that $0 \le x \le n$. Suppose the population size N and the number of successes M has the property that

$$\lim_{N\to\infty} \frac{M}{N} = p.$$

Show that

$$\lim_{N\to\infty} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \binom{n}{x} p^x (1-p)^{n-x}.$$

Hint: Use the exercise above on each binomial coefficient, then pass to a limit as $N \to \infty$ keeping in mind that $M/N \to p$, $(N - M)/N \to 1 - p$ as $N \to \infty$, and some terms approach 1 as $N \to \infty$. This is an exercise in good book-keeping.

The next discrete distribution we discuss is the Poisson(λ) distribution, but we first need to recall a couple of calculus facts.

Calculus fact: MacLaurin series for e^u.
For any (real) u,

$$e^u = \sum_{k=0}^{\infty} \frac{u^k}{k!} = 1 + u + \frac{u^2}{2!} + \frac{u^3}{3!} + \frac{u^4}{4!} + \cdots .$$

Calculus fact: limit representations for e^u.
For any (real) u,

$$e^u = \lim_{n\to\infty} \left(1 + \frac{u}{n}\right)^n = \lim_{n\to\infty} \left(1 + \frac{u}{n} + o(\frac{1}{n})\right)^n .$$

4. The Poisson(λ) distribution.

This is the distribution of a discrete rv X that counts the number of "events" that happen in a fixed amount of exposure (e.g., amount of time, or amount of space) that roughly satisfy the following criteria:

randomness:
the "events" occur randomly/independently throughout the exposure.
constancy:
the "events" happen at a constant rate $\lambda > 0$ throughout the exposure.
no-clumping:
the chance of two or more "events" happening at the same point is negligible.

In this case we write $X \sim$ Poisson(λ). We'll show that the ***pmf of the Poisson***(λ) is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for } x = 0, 1, 2, 3, \ldots .$$

To understand the "no-clumping" criterion, represent the unit of exposure as the unit interval $[0, 1]$. For $t \in [0, 1]$, let $X(t)$ be the number of events in $(0, t]$, $X(0) = 0$. Then $X(t) - X(t - \frac{1}{n})$ represents the number of events in $(t - \frac{1}{n}, t]$. Now, rigorously, we understand "no-clumping" means: for each t, $P(X(t) - X(t - \frac{1}{n}) \geq 2) = o(1)$ as $n \to \infty$.

Remark and discussion.
The following random variables can be modeled well by a Poisson(λ) distribution.
• number of babies born in a day at a given hospital (say, $\lambda = 10$ babies/day).
• number of category-5 hurricanes making landfall in a year (say, $\lambda = 0.4$ hurricanes/year).
• number of defects in a square yard of cloth (say, $\lambda = 2$ defects/yard^2).

Consider the example of babies being born in a hospital. The Poisson "events" that we are counting are births of babies. In any given day, say starting from the stroke of midnight, as time continually moves we will see a first event: the first baby is born. Then a second baby is born at some time *after* the first, and so on.

The randomness assumption says the birth of a baby will not influence the probability of another birth - and this assumption seems plausible in this example.

The constancy assumption says the rate that babies are born in the day is constant throughout the day, i.e., that the Poisson events should not occur at, say, higher rates in certain subintervals of time than others within the day. This also seems plausible. An example where we might expect that constancy is not satisfied is the case of cars passing through a particular intersection in a day. In this case, one might suspect that the volume of cars is higher say in the mid-morning and late afternoon traffic and have rather light volume in the early morning hours.

The no-clumping assumption for baby births might be challenged, but even twins are not born simultaneously and this appears somewhat plausible.

Example.
Show that $P(X = x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, 3, \ldots$ is, indeed, a pmf.

Since $\lambda > 0$, $\dfrac{e^{-\lambda}\lambda^x}{x!} > 0$ for all $x \in \mathbb{N}$, and

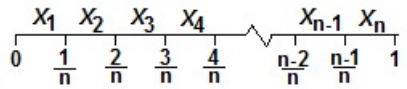
$$\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1,$$

shows the masses sum to 1, so *is* a pmf. Notice in the second equality we recognize the MacLaurin expansion of e^{λ}.

How does the Poisson(λ) pmf come about?

Let's consider the example of babies being born in a hospital on a given day. We represent the exposure (1 day) as the unit interval $0 \leq x \leq 1$. This a continuous interval starting at time 0 (the beginning of the day) and ending at time 1 (the end of the day). We should envision babies being born in this interval at a constant rate but entirely at random – meaning that a baby being born at one instant will not influence the probability that a baby is born at another instant – and that two or more babies cannot be born at the same instant. The random variable X will count the number of babies born in the interval $[0, 1]$.

The idea is to "chop up" the unit interval into a large number of subintervals of equal length, say n pieces of length $\frac{1}{n}$ each, and count the number of babies born in each subinterval, say X_j counts the number of babies born in the interval $(\frac{j-1}{n}, \frac{j}{n}]$. See figure.



According to our randomness assumption, $X_1, X_2, \ldots, X_n$ should be independent rvs. Since each subinterval is the same length, the constancy assumption implies that each X_i should have the same probability distribution. Now, if n is *really* large, then because of the no-clumping assumption, for any j, $P(X_j \geq 2) \approx 0$. This means for large n, each X_j is (approximately) Bernoulli (since each X_j can only take the values 0 and 1, neglecting the case of 2 or more). Lastly, since λ represents the expected number of babies born in the entire interval $[0, 1]$, the constancy assumption says we should expect $\frac{\lambda}{n}$ babies in each subinterval. Putting this all together we have, for each n, the sequence $X_1, X_2, \ldots, X_n$ are independent Bernoulli($\frac{\lambda}{n}$) rvs and their sum

$$S = S_n = X_1 + X_2 + \cdots + X_n \sim \text{binom}(n, \frac{\lambda}{n}).$$

Letting n tend to infinity, we should get our Poisson(λ). Let's see:
Fix $x \in \mathbb{N}$, take $n > x$ with the goal that $n \to \infty$ eventually.

$$\begin{aligned}
P(S_n = x) &= \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \frac{n(n-1)(n-2)\cdots(n-(x-1))}{x!}\left(\frac{\lambda^x}{n^x}\right)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!}\cdot\frac{n}{n}\cdot\frac{(n-1)}{n}\cdot\frac{(n-2)}{n}\cdots\frac{(n-(x-1))}{n}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!}\cdot\underbrace{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)}_{\to 1 \text{ as } n\to\infty \text{ since } x \text{ is fixed}}\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda} \text{ as } n\to\infty}\underbrace{\left(1-\frac{\lambda}{n}\right)^{-x}}_{\to 1 \text{ as } n\to\infty} \\
&\longrightarrow \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{as } n \to \infty.
\end{aligned}$$

Example.
Suppose the number X of category-5 hurricanes that make landfall in a year has a Poisson(0.4) distribution. What is the probability that at least one category-5 hurricane makes landfall next year? How about the next 5 years?

In the next year, we expect $\lambda = 0.4$, so

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-0.4}\frac{(0.4)^0}{0!} = 1 - e^{-0.4} \approx 0.3297.$$

In the next 5 years, we expect $\lambda = 5(0.4) = 2$ hurricanes (notice the change in exposure affected the value of λ chosen).
$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-2} \approx 0.8647.$

Remark.
Page 108 gives a nice connection between the Poisson and binomial distributions: the Poisson(λ) distribution can be viewed as a limit as $n \to \infty$ of the binomial($n, \frac{\lambda}{n}$) and we should expect the Poisson(λ) pmf to be close to the binomial($n, \frac{\lambda}{n}$) pmf for large n; i.e., the rv counts the number of successes in a large number of independent Bernoulli trials where successes are "*rare*" is *approximately* Poisson.

5. The geometric(p) distribution.

Suppose X is the discrete rv that returns the ***trial of the first success*** in a sequence of independent Bernoulli(p) trials. Then, we write $X \sim$ geometric(p) and

$$p(x) = P(X = x) = p(1-p)^{x-1} \quad \text{for } x = 1, 2, 3, \ldots.$$

(Why?)

Certainly, $P(X = 1) = p$ since the event $(X = 1)$ is the same as the event that we obtain a success in one trial; and, the event $(X = 2)$ is the same as the event that our first success follows a single failure, and the independence of the trials implies $P(X = 2) = (1-p)p$. More generally, if $x > 1$, then the event $(X = x)$ is the event that the first success follows $x - 1$ failures, and independence of the trials implies

$$P(X = x) = \underbrace{(1-p)(1-p)\cdots(1-p)}_{x-1 \text{ factors}} p = p(1-p)^{x-1}.$$

Remark.
An alternative way of defining a geometric(p) distribution is to have the rv count the numbers of failures before the first success. In this way, this interpretation of the geometric rv, say Y, would take values $0, 1, 2, \ldots$ and

$$P(Y = y) = p(1-p)^{y} \quad \text{for } y = 0, 1, 2, \ldots$$

would be the corresponding pmf. Either of these distributions are geometric(p) distributions, we can choose which to use from the context of the problem. Of course, they are related by $X = 1 + Y$.

Interesting examples of the geometric distribution.

• Roll a pair of fair dice and stop when you see a *double six* for the first time. The trial X where you stop rolling is a geometric($\frac{1}{36}$).

• You are interviewing people one at a time and we label the people 1, 2, 3, and so on. The first person X to have your birthday has a geometric($\frac{1}{365}$) distribution.

• Fred has a 60% chance of hitting the bulls-eye on any given dart throw (when he's aiming for it). The number of dart throws Fred needs to hit the bulls-eye for the first time has a geometric(0.6) distribution.

Calculus fact: Sums of a geometric series.
For any constants A and r, the sequence $A, Ar, Ar^2, Ar^3, \ldots$ is called a ***geometric progression***. The value r is called the ***geometric ratio***. When $|r| < 1$ the ***sum of the geometric series*** is

$$\sum_{k=m}^{\infty} Ar^k = \frac{Ar^m}{1-r}.$$

Advice:
Become familiar with recognizing when the terms of a sum/series form a geometric progression, and ***memorize*** the formulas for the sums. A good mnemonic for remembering the formula for a geometric series is "it's the first term in the series divided by 1 minus the geometric ratio".

Example.
For $0 < p < 1$, show that $P(X = x) = p(1-p)^{x-1}$ for $x = 1, 2, 3, \ldots$ is a pmf.

Clearly, $p(1-p)^{x-1} > 0$ for all $x \geq 1$. Moreover, by the calculus fact above

$$\sum_{x=1}^{\infty} p(1-p)^{x-1} = \frac{p(1-p)^{1-1}}{1-(1-p)} = 1.$$

Examples.
You plan to roll a pair of 6-sided dice until you get *double sixes*. What's the probability you succeed by the 3rd roll?

$P(X \leq 3) = \frac{1}{36} + \frac{1}{36}\left(\frac{35}{36}\right) + \frac{1}{36} \cdot \left(\frac{35}{36}\right)^2 \approx 0.08104.$

You are interviewing people one at a time. What's the probability that in the first 400 interviews no one had your birthday? This means that the first person to have your birthday is *after* 400.

$P(X > 400) = \sum_{x=401}^{\infty} \frac{1}{365}\left(\frac{364}{365}\right)^{x-1} = \frac{\frac{1}{365}\left(\frac{364}{365}\right)^{400}}{1-\left(1-\frac{1}{365}\right)} = \left(\frac{364}{365}\right)^{400} \approx 0.33374.$

What's the probability that the first person to have your birthday was interviewed between person 100 and person 200 (inclusive)?

$P(100 \leq X \leq 200) = P(X > 99) - P(X > 200) = \left(\frac{364}{365}\right)^{99} - \left(\frac{364}{365}\right)^{200} \approx 0.18445.$

Fred has a 60% chance of hitting the bulls-eye on any given dart throw (when he's aiming for it). Compute the probability that Fred's first bulls-eye happens on the 3rd throw.

$P(X = 3) = 0.6(0.4)^2 = 0.096.$

This next distribution generalizes the geometric(p) distribution.

6. The negative binomial(r, p) distribution.

If X represents the ***trial of the rth success*** in a sequence of independent Bernoulli(p) random variables, then X is said to have a negative binomial(r, p) distribution, which we write as $X \sim$ neg.binom(r, p), and the pmf of X is

$$p(x) := P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad \text{for } x = r, r+1, r+2, \ldots.$$

(Why?)

Fix an $r \geq 1$. Clearly, we need at least $x = r$ trials to see the rth success, so x needs to be at least r. Now if we fix $x \geq r$, then the event $(X = x)$ means the rth success happens on the trial x, which tacitly implies there must be $r - 1$ successes in the $x - 1$ previous trials. Thus, we have

$$(X = x) = \{r - 1 \text{ successes in first } x - 1 \text{ trials}\} \cap \{\text{success on trial } x\},$$

where the two events in the intersection are independent. Consequently,

$$P(X = x) = \underbrace{P(r - 1 \text{ successes in first } x - 1 \text{ trials})}_{=\binom{x-1}{r-1} p^{r-1}(1-p)^{x-1-(r-1)}} \times \underbrace{P(\text{success on trial } x)}_{=p}.$$

Remark.
Notice that when $r = 1$, this is just the geometric(p) distribution.

Just as in the remark on page 110 there is an alternate way of defining a negative binomial rv. Define $Y = X - r$ to be the number of failures before the rth success. In this way, this rv Y would take values $0, 1, 2, \ldots$ and the resulting pmf would be

$$P(Y = y) = \binom{y + r - 1}{y} p^r (1-p)^y \quad \text{for } y = 0, 1, 2, \ldots.$$

Which form of the negative binomial we want to choose should be clear from context, but either X (the trial of the rth success) or Y (the number of failures before the rth success in independent Bernoulli(p) trials) are neg.binom(r, p) rvs.

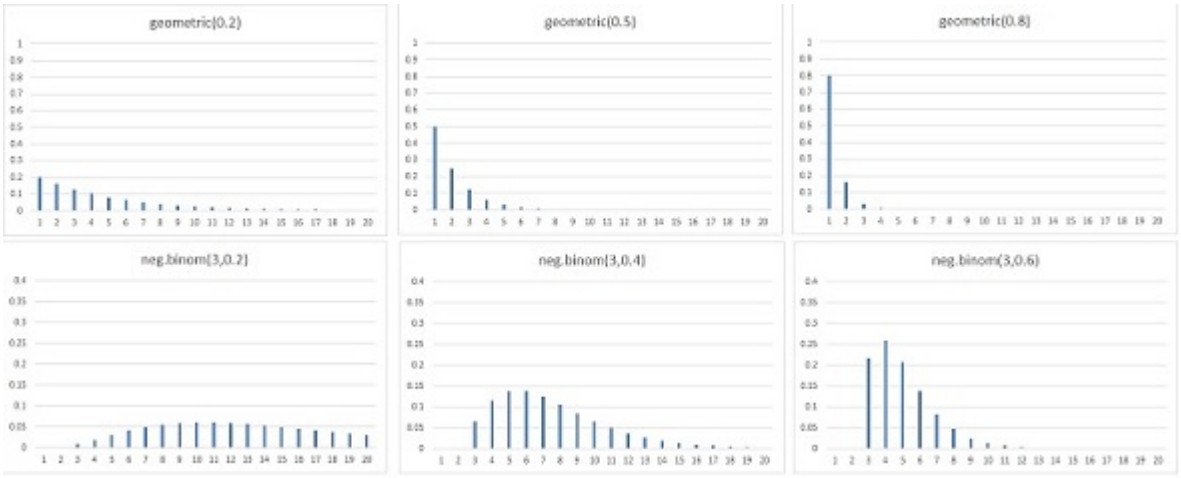


Figure. Some plots of negative binomial pmfs.

Example. We roll a fair (balanced) 6-sided die repeatedly. Find the probability that the 4th '6' occurs at the 10th roll.

Here, if X is the roll on which the 4th '6' appears, then $X \sim \text{neg.binom}(4, \frac{1}{6})$.
$P(X = 10) = \binom{9}{3}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^6 \approx 0.0217$.

Example.
What's the probability that see your third 6 by the 5th roll?

If X is the roll of the third 6, then $X \sim \text{neg.binom}(3, \frac{1}{6})$ and

$$\begin{aligned} P(\underbrace{X \leq 5}_{3,4,5}) &= P(X=3) + P(X=4) + P(X=5) \\ &= \binom{2}{2}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^0 + \binom{3}{2}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1 + \binom{4}{2}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 \\ &\approx 0.0355. \end{aligned}$$

Challenging exercise.
Show that the $p(x)$ above is a pmf; In particular, use mathematical induction to prove

$$\sum_{x=r}^{\infty} \binom{x-1}{r-1} p^r (1-p)^{x-r} = 1.$$

When $r = 1$ this is just the geometric(p) which we already showed sums to 1 (see page 111). I'll now do the case $r = 2$. I'll need the following ***combinatorial identity***:
For integer $n > 1$ and $0 < r \leq n$,

$$\binom{n}{r} = \binom{r-1}{r-1} + \binom{r}{r-1} + \binom{r+1}{r-1} + \cdots + \binom{n-2}{r-1} + \binom{n-1}{r-1} = \sum_{j=r-1}^{n-1} \binom{j}{r-1}.$$

Try to construct a proof by mathematical induction for this identity by appealing to the ***Pascal identity*** repeatedly on $\binom{n}{r}$.

This identity allows us to represent, for instance,

$$\begin{aligned} \binom{6}{3} &= \binom{2}{2} + \binom{3}{2} + \binom{4}{2} + \binom{5}{2} \\ &= 1 + 3 + 6 + 10 = 20, \end{aligned}$$

and

$$\begin{aligned} \binom{n+1}{2} &= \binom{1}{1} + \binom{2}{1} + \binom{3}{1} + \cdots + \binom{n}{1} \\ &= 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}. \end{aligned}$$

To show that

$$\sum_{x=2}^{\infty} \binom{x-1}{1} p^2 (1-p)^{x-2} = 1,$$

we use the above combinatorial identity to represent $\binom{x-1}{1} = \sum_{j=0}^{x-2} \binom{j}{0} = \sum_{j=0}^{x-2} 1.$

Therefore,

$$\begin{aligned}\sum_{x=2}^{\infty}\binom{x-1}{1}p^2(1-p)^{x-2} &= \sum_{x=2}^{\infty}\sum_{j=0}^{x-2}p^2(1-p)^{x-2}\\ &= \sum_{j=0}^{\infty}\sum_{x=j+2}^{\infty}p^2(1-p)^{x-2}\\ &= \sum_{j=0}^{\infty}\frac{p^2(1-p)^j}{1-(1-p)}\\ &= \sum_{j=0}^{\infty}p(1-p)^j = 1,\end{aligned}$$

where, in the second equality, we changed the order of summation. Now use mathematical induction to complete the proof for general $r \geq 1$.

Analyzing the special case of a neg.binom$(2, p)$.
In this case X is the trial of the second success, which means there is a first success that happened before it. So, *fix* an $x \geq 2$ and consider the event, $(X = x)$, that the second success happens at trial x. Since a first success must happen before x, we can decompose this event as the mutually exclusive union

$$(X = x) = \bigcup_{k=1}^{x-1}(X_1 = k, X = x),$$

where X_1 is the trial of the first success (a geometric(p) rv). Once this first success happens at $X_1 = k$, the trials performed thereafter are independent of the trials before. So, it's just like starting anew after the first success and getting a second geometric(p) rv, say X_2. See the figure below. X_2 represents the number of trials that *elapsed* after the first success until the next success.

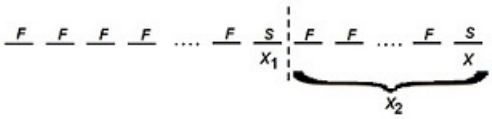


Figure. A particular realization of the neg.binom$(2, p)$.

Therefore,

$$(X = x) = \bigcup_{k=1}^{x-1}(X_1 = k, X - X_1 = x - k) = \bigcup_{k=1}^{x-1}(X_1 = k, X_2 = x - k),$$

and

$$\begin{aligned} P(X=x) &= \sum_{k=1}^{x-1} P(X_1=k, X_2=x-k) \\ &= \sum_{k=1}^{x-1} P(X_1=k)P(X_2=x-k) \\ &= \sum_{k=1}^{x-1} p(1-p)^{k-1} \cdot p(1-p)^{x-k-1} \\ &= \sum_{k=1}^{x-1} p^2(1-p)^{x-2} = \binom{x-1}{1} p^2(1-p)^{x-2}. \end{aligned}$$

Remark.
The analysis done above suggests that the neg.binom$(2,p)$ is the sum of two independent geometric(p) random variables. In fact, later in this course, $X \sim$ neg.binom(r,p) can be written as $X = X_1 + X_2 + \cdots + X_r$ where $X_1, X_2, \ldots, X_r$ are *independent* geometric(p) random variables.

Expected values of discrete random variables.
If X is a discrete rv with pmf $p_X(x) = P(X = x)$, then we define

$$E(X) = \sum_x x\, p_X(x) = \sum_x x\, P(X = x).$$

The sum is taken over ***all*** possible values of the rv X. $E(X)$ is called any of the following: ***expected value of*** X, ***expectation (value) of*** X, ***the mean (value) of*** X.

Advice:
Try to remember $E(X)$ as a real number representing the weighted average of the values of X where the weights are just the respective probability masses.

Example and Discussion.
Suppose X has the pmf $P(X = -3) = .3,\ P(X = 2) = .6,\ P(X = 7) = .1$ Then

$$E(X) = -3(.3) + 2(.6) + 7(.1) = 1.$$

Statisticians usually think of a random variable as representing a (typically infinite) population of real values, and, from this viewpoint, $E(X)$ – often denoted μ or μ_X – is called the ***(population) mean*** or ***population average***. Here's the intuition. Let's first imagine that X represents a *finite* "population", say, of 10 objects: 3 of these objects are -3, 6 of these objects are 2, and 1 of them is 7. Then, since this population is finite, our population average μ is just:

$$\begin{aligned} \frac{-3 + -3 + -3 + 2 + 2 + 2 + 2 + 2 + 2 + 7}{10} &= -3\left(\frac{3}{10}\right) + 2\left(\frac{6}{10}\right) + 7\left(\frac{1}{10}\right) \\ &= E(X) = 1. \end{aligned}$$

We see the "straight average" of the population is just the expected value of the random variable representing it. This intuition extends naturally if the population had consisted of an infinite number of objects 30% of which are -3, 60% of which are 2, and 10% of which are 7 by defining μ to be $E(X)$. I'll mention that the expected value does ***not*** need to be a possible value of the rv as this example clearly demonstrates.

If we think of the pmf as how the (probability) mass is distributed to the values of the discrete rv, then the expected value can be interpreted as the center of (probability) mass - in physics, this is called the ***first moment of inertia***.

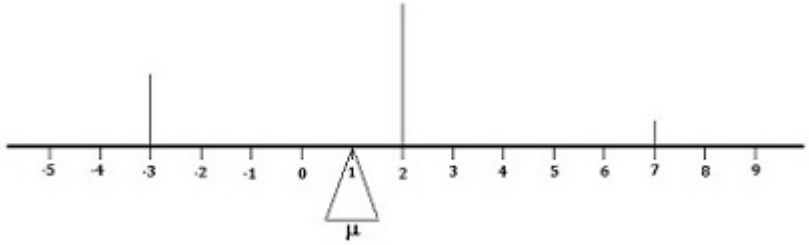


Figure. We balance the see-saw at μ.

Example.
Let's compute the ***mean of a Bernoulli***(p).
Recall if $X \sim$ Bernoulli(p), then $P(X=1)=p$ and $P(X=0)=1-p$). Therefore,

$$E(X) = \sum_{x\in\{0,1\}} x\,P(X=x) = 0\cdot P(X=0) + 1\cdot P(X=1) = p.$$

Example.
Let's show the ***mean of a binomial***(n,p) is np.
We will first compute $E(X)$ using the definition.

$$\begin{aligned}
E(X) &= \sum_{x=0}^{n} x\,P(X=x) \\
&= \sum_{x=\mathbf{1}}^{n} x\cdot\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \quad (x=0 \text{ will not contribute to } E(X)) \\
&= \sum_{x=1}^{n}\frac{n!}{(x-1)!(n-x)!}p^x(1-p)^{n-x} \\
&= np\sum_{x=1}^{n}\underbrace{\frac{(n-1)!}{(x-1)!(n-x)!}}_{=\binom{n-1}{x-1}}p^{x-1}(1-p)^{n-1-(x-1)} \quad (\text{change of variable } u=x-1) \\
&= np\underbrace{\sum_{u=0}^{n-1}\binom{n-1}{u}p^u(1-p)^{n-1-u}}_{=\,1 \text{ summing binom}(n-1,p) \text{ over its support}} \\
&= np.
\end{aligned}$$

If you understood the calculation for the mean of a binomial(n,p) above, then do this

Exercise.
Compute the mean of a Poisson(λ). ***Now!***

Example.
Let's show the ***mean of a geometric***(p) ***is*** $\dfrac{1}{p}$.

I'll actually do this calculation 2 ways.

solution #1:

$$E(X) = \sum_{x=1}^{\infty} x\, p(1-p)^{x-1} = p + 2p(1-p) + 3p(1-p)^2 + 4p(1-p)^3 + \cdots . \qquad (3)$$

Assume this series is convergent for now; multiply equation (3) through by $1-p$:

$$(1-p)E(X) = \sum_{x=1}^{\infty} x\, p(1-p)^{x} = p(1-p) + 2p(1-p)^2 + 3p(1-p)^3 + 4p(1-p)^5 + \cdots . \quad (4)$$

Now write equations (3) and (4) aligned by like terms:

$$\begin{array}{rcccccccccc} E(X) & = & p & + & 2p(1-p) & + & 3p(1-p)^2 & + & 4p(1-p)^3 & + & \cdots \\ (1-p)E(X) & = & & & p(1-p) & + & 2p(1-p)^2 & + & 3p(1-p)^3 & + & \cdots \end{array}$$

I want to subtract equation (4) from (3) which is allowed because (by assumption) both these series converge: the difference of the series is the series of the differences. So, we can subtract term by term. On the left we get

$$E(X) - (1-p)E(X) = pE(X).$$

On the right we get

$$p + p(1-p) + p(1-p)^2 + p(1-p)^3 + \cdots = 1$$

since this is just a geometric series. Thus,

$$pE(X) = 1 \implies E(X) = \frac{1}{p}.$$

solution #2:
If we set $q := 1-p$, then

$$\begin{array}{rcll} E(X) & = & (1-q)\sum_{x=1}^{\infty} x\, q^{x-1} & \\ & = & (1-q)\sum_{x=1}^{\infty} \frac{d}{dq}\left(q^x\right) & \text{(power rule)} \\ & = & (1-q)\cdot \frac{d}{dq}\sum_{x=1}^{\infty} q^x & \text{($\frac{d}{dq}$ is a linear operator)} \\ & = & (1-q)\cdot \frac{d}{dq}\left(\frac{q}{1-q}\right) & \text{(sum of a geometric series)} \\ & = & (1-q)\cdot \frac{1}{(1-q)^2} = \frac{1}{1-q} = \frac{1}{p}. & \text{(using the derivative quotient rule)} \end{array}$$

Example.
If $X \sim$ neg.binom.$(2, p)$, compute $E(X)$.

solution #1:
By brute force use of definition

$$\begin{aligned} E(X) &= \sum_{x=2}^{\infty} x\,(x-1)p^2(1-p)^{x-2} \\ &= 2(1)p^2 + 3(2)p^2(1-p) + 4(3)p^2(1-p)^2 + 5(4)p^2(1-p)^3 + \cdots \\ (1-p)E(X) &= \qquad +2(1)p^2(1-p) + 3(2)p^2(1-p)^2 + 4(3)p^2(1-p)^3 + \cdots \end{aligned}$$

Subtracting

$$pE(X) = 2(1)p^2 + 2(2)p^2(1-p) + 2(3)p^2(1-p)^2 + 2(4)p^2(1-p)^3 + \cdots .$$

Multiplying through by $1-p$ in this last equation we obtain

$$(1-p)pE(X) = \qquad +2(1)p^2(1-p) + 2(2)p^2(1-p)^2 + 2(3)p^2(1-p)^3 + \cdots ,$$

and, again, subtracting

$$p^2E(X) = 2p^2 + 2p^2(1-p) + 2p^2(1-p)^2 + 2p^2(1-p)^3 + \cdots = \frac{2p^2}{1-(1-p)} = 2p$$

and $E(X) = \frac{2}{p}$.

solution #2:
From the remark on page 116 we learned if $X \sim$ neg.binom$(2, p)$, then $X = X_1 + X_2$, where X_1 and X_2 are geometric(p) random variables, in fact, *independent* geometric(p)'s. But then, by linearity of expectation, $E(X) = E(X_1+X_2) = E(X_1)+E(X_2) = \frac{1}{p}+\frac{1}{p} = \frac{2}{p}$.

Properties of expected value.

Following properties are true for rvs in general not just discrete rvs assuming the expected values exist.

Property #1:
For any constants a and b and random variable X:

$$E(aX + b) = aE(X) + b.$$

Proof. Assuming X is discrete, and using LOTUS,

$$\begin{aligned} E(aX+b) &= \sum_x (ax+b)P(X=x) \\ &= \sum_x \left(ax\,P(X=x) + b\,P(X=x)\right) \\ &= \sum_x ax\,P(X=x) + \sum_x b\,P(X=x) \\ &= \underbrace{a\sum_x x\,P(X=x)}_{=E(X)} + \underbrace{b\sum_x P(X=x)}_{=1}. \end{aligned}$$

□

Two special cases:

If $a = 0$ and b is constant, then $E(b) = b$. This says the expected value of a constant is the constant.

If a is constant and $b = 0$, then $E(aX) = aE(X)$. This says constant factors can be taken out of the expectation.

Property #2: (**Linearity of expectation**)
If the expected values of the rvs $X_1, X_2, \ldots, X_n$ exist, then

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i).$$

We will give a proof of this later in the course.

Expected values: existence and nonexistence

This discussion pertains to ***all*** random variables - not just discrete ones. For a discrete random variable X we defined

$$E(X) = \sum_x x\, P(X = x).$$

When X is finitely supported there is no ambiguity in this definition - namely, it is just a finite sum of finite numbers and thus will be finite. Now suppose the random variable is not finitely supported, i.e., can take on arbitrarily large positive and/or negative values, then our definition of expected value is no longer a finite sum and, in fact, becomes an *infinite series*. So care must be taken in interpreting when the expected value exists!

First suppose X is nonnegative but not finitely supported. Then, since $X \geq 0$ it is clear that $E(X) \geq 0$. We will say the expected value of X exists if $0 \leq E(X) < \infty$, i.e., the infinite series converges, and in this case the expected value is bounded above and below by finite numbers; otherwise, $E(X) = +\infty$ and we say the expected value is infinite. Similarly, if $X \leq 0$ but not finitely supported, then $E(X) \leq 0$. In this case, if $E(X) > -\infty$, then $-\infty < E(X) \leq 0$ and we will say the expected value of X exists and is finite since it is, again, bounded above and below by finite numbers; otherwise, $E(X) = -\infty$ and we say the expected value is infinite. FYI: If $X \leq 0$, then $-X \geq 0$ and everything written in these last two sentences falls into first case described in this paragraph - so handling the negative case was redundant.

What if X is not finitely supported but can take on both positive and negative values? In this case, probabilists have agreed to write X as the difference between its positive and negative parts. Let's define ***the positive part of*** X as

$$X^+ = \max\{X, 0\}$$

and ***the negative part of*** X as

$$X^- = \max\{-X, 0\}.$$

X^+ and X^- are *nonnegative* random variables and

$$X = X^+ - X^-.$$

We then say that the the expected value ***exists and is finite*** and label if μ if both $E(X^+)$ and $E(X^-)$ exist and are finite. By the way, we point out that $|X| = X^+ + X^-$ and therefore, $E(X)$ exists and is finite if and only if $E(|X|)$ exists and is finite!

Thus, $E(X)$ can only exist and be finite when both $E(X^+)$ and $E(X^-)$ are both finite. $E(X)$ will not exist if $E(X^+)$ and $E(X^-)$ are *both* infinite.

Calculus fact:* p*-series, and the harmonic series diverges.
The series

$$\sum_{n=1}^{\infty} \frac{1}{n^p}$$

is called a p***-series***.

When $0 < p \leq 1$, this series diverges, i.e., it is infinite.

When $p > 1$, the series converges.

In the special case $p = 1$, we call the series

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

the ***harmonic series*** which diverges. (Can you give a proof of this?)

Example. (A random variable with infinite expected value)
The ***Riemann Zeta function*** is defined as $\zeta(p) = \sum_{n=1}^{\infty} \frac{1}{n^p}$. We consider this function only on the domain $\{p \in \mathbb{R} : p > 1\}$ so that for p in this domain, $0 < \zeta(p) < \infty$. It follows that

$$\frac{\frac{1}{n^p}}{\zeta(p)} \quad \text{for } n = 1, 2, 3, \ldots$$

is a probability mass function and is often called the Zeta(p) distribution.

Now, consider a (discrete) random variable X having the Zeta(2) distribution. It can be shown that $\zeta(2) = \frac{\pi^2}{6}$ but this is not important. Compute $E(X)$.

$$E(X) = \sum_{n=1}^{\infty} n\, P(X = n) = \sum_{n=1}^{\infty} n \cdot \frac{\frac{1}{n^2}}{\zeta(2)} = \frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n} = +\infty.$$

In a sense the reason this random variable has infinite expected value is because the tail probabilities are too large, they don't go to zero fast enough to make the series converge.

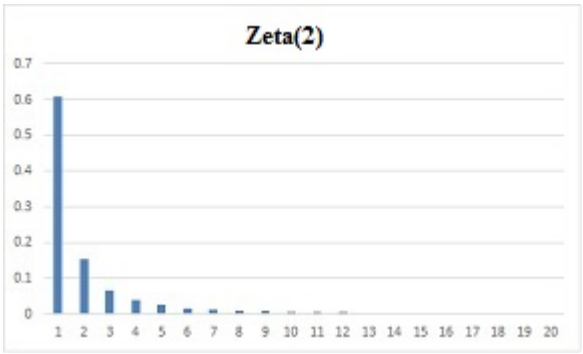


Figure. A plot of the Zeta(2) pmf.

Example. (An example of a distribution whose expected value does not exist)
Consider a discrete random variable X with the pmf

$$P(X = n) = \frac{3}{\pi^2} \cdot \frac{1}{n^2} \quad \text{for } n = \pm 1, \pm 2, \pm 3, \ldots.$$

Notice, in this example, that X takes on both arbitrarily large positive and negative values. Moreover,

$$E(X^+) = E(X^-) = \frac{3}{\pi^2} \sum_{n=1}^{\infty} n \cdot \frac{1}{n^2} = +\infty.$$

Consequently, since both positive and negative parts of X have infinite expected values, $E(X)$ ***does not exist***.

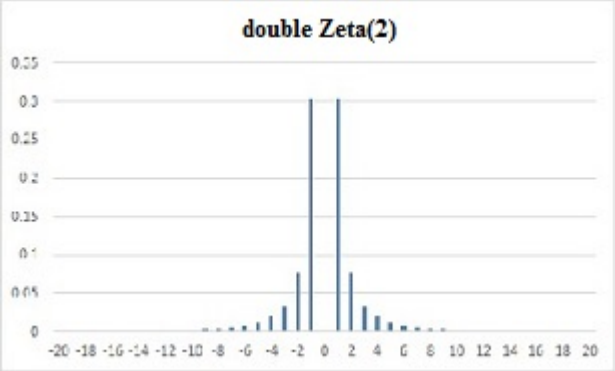


Figure. plot of the double Zeta(2) distribution.

Remark.
It might seem strange at first to say that the expected value in the last example does not exist since the distribution is symmetric about $X = 0$ and one might want to believe that $E(X)$ is zero. But, *by definition*, we are saying this expected value does not exist (because both the positive and negative parts have infinite expectations). The calculus reason is that we are demanding our infinite series to be *absolutely* convergent and not just conditionally convergent.

The infinite series $E(X)$ is a sum of positive and negative terms. An infinite series $\sum_n a_n$ is called ***absolutely convergent*** if the series $\sum_n |a_n|$ of the absolute values of each of these terms also converges. If the series $\sum_n a_n$ converges but is not absolutely convergent, then it is ***conditionally convergent***. In the above example, $E(X)$ is only conditionally convergent and not absolutely convergent.

The following comparison of facts regarding conditionally convergent and absolutely convergent series is helpful. No matter how the terms of an absolutely convergent series are rearranged, the sum will always be the *same*. However, if we have a conditionally convergent series that is not absolutely convergent (i.e., the sum of the absolute values is infinite), then the terms of the series can be rearranged to sum at any value we desire. This non-uniqueness is the issue we want to avoid in our definition of expected value.

Sometimes we have or know the distribution of a random variable X but we want to compute the expected value of some function of X instead. For example, suppose we have a very good model for the discrete random variable X representing the quantity produced, i.e., we know the pmf, say, $P(X = x)$. We have a *cost function* associated with producing the quantity x, namely,

$$C(x) = 0.1x^3 - 2x^2 + 60x + 200.$$

Then $C(X)$ is a random variable, and we may like to compute the expected cost $E[C(X)]$.

The following theorem can be very helpful in situations like this.

Law of the Unconscious Statistician (LOTUS).
Suppose X is a discrete random variable and $g = g(x)$ is a function. Then

$$E[g(X)] = \sum_x g(x)\, P(X = x),$$

assuming this expected value exists.

Proof. Set $Y = g(X)$ and let y be a value of the random variable $Y = g(X)$. Then, by definition,

$$\begin{aligned} E[g(X)] &= \sum_y y\, P[g(X) = y] \\ &= \sum_y y \sum_{x:g(x)=y} P(X = x) \\ &= \sum_y \sum_{x:g(x)=y} y\, P(X = x) \\ &= \sum_y \sum_{x:g(x)=y} g(x)\, P(X = x) \\ &= \sum_x g(x)\, P(X = x). \end{aligned}$$

□

Example. Suppose X has the pmf:

x	-2	-1	0	1	2	5
$P(X = x)$	0.4	0.2	0.1	0.1	0.1	0.1

Compute $E(X^2)$.

Using LOTUS,

$$\begin{aligned} E(X^2) &= \sum_{x\in\{-2,-1,0,1,2,5\}} x^2\, P(X = x) \\ &= (-2)^2\,(0.4) + (-1)^2\,(0.2) + 0^2\,(0.1) + 1^2\,(0.1) + 2^2\,(0.1) + 5^2\,(0.1) \\ &= 1.6 + 0.2 + 0 + 0.1 + 0.3 + 2.5 = \mathbf{4.8}. \end{aligned}$$

Remark.
Without LOTUS how would we compute $E(X^2)$? We'd have to use the definition which says to take the weighted average of the values of the rv $Y = X^2$ weighted against the pmf of Y. Therefore, we'd need to derive the pmf of Y since it isn't directly given to us. From the last example, the possible values of $Y = X^2$ are $y = 0, 1, 4$, and 25. Moreover,

$P(Y = 0) = P(X = 0) = 0.1$
$P(Y = 1) = P(X = 1 \cup X = -1) = P(X = 1) + P(X = -1) = 0.1 + 0.2 = 0.3$
$P(Y = 4) = P(X = 2 \cup X = -2) = P(X = 2) + P(X = -2) = 0.1 + 0.4 = 0.5$
$P(Y = 25) = P(X = 5) = 0.1$

is the pmf for Y. Therefore, we'd compute

$$E(Y) = 0 \times 0.1 + 1 \times 0.3 + 4 \times 0.5 + 25 \times 0.1 = 4.8.$$

Of course, we get the *same* answer, but, LOTUS allowed us to bypass having to compute the pmf of X^2 and this sure saved us a lot of time!

Moments of a random variable/distribution.
Let X be a random variable. We define

$$\mu_k := E(X^k)$$

to be the k***th moment of*** X or the k***th moment of the distribution of*** X.

Remark.
Notice that $\mu_0 = 1$ for any random variable and, therefore, when we talk about the moments of a random variable we typically mean for $k = 1, 2, 3, \ldots$. Also, when $k = 1$, the ***first moment*** is just μ, i.e., the ***mean*** of X. We'll see that the moments of a random variable reveal information about the underlying distribution of the random variable. In a sense the more moments we know the more information we have about the probability distribution of the rv.

Example.
Let $X \sim$ Bernoulli(p).
Use LOTUS to compute the kth moment of the Bernoulli(p), i.e., $E(X^k)$ for integer $k \geq 1$.

$E(X^k) = 0^k P(X = 0) + 1^k P(X = 1) = p$ for every $k \geq 1$.

Variance.
Let X be a random variable having a finite mean μ. Then $X - \mu$ is the deviation X makes from its mean, and $(X-\mu)^2$ is the squared deviation X makes from its mean. We define

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

Remark.
The variance of X measures how spread out the values of the random variables are from its mean. Clearly, $Var(X) \geq 0$ always. The larger the variance is, the more spread out the values are from its mean – the values of X can fluctuate wildly from the mean. The closer the variance is to 0, the less spread out the values are from the mean – the values do not fluctuate much from its mean.

If we have two rvs X and Y and we know $Var(X) < Var(Y)$, then the values in the distribution for Y are more spread out than those for the distribution X. See the figure below.

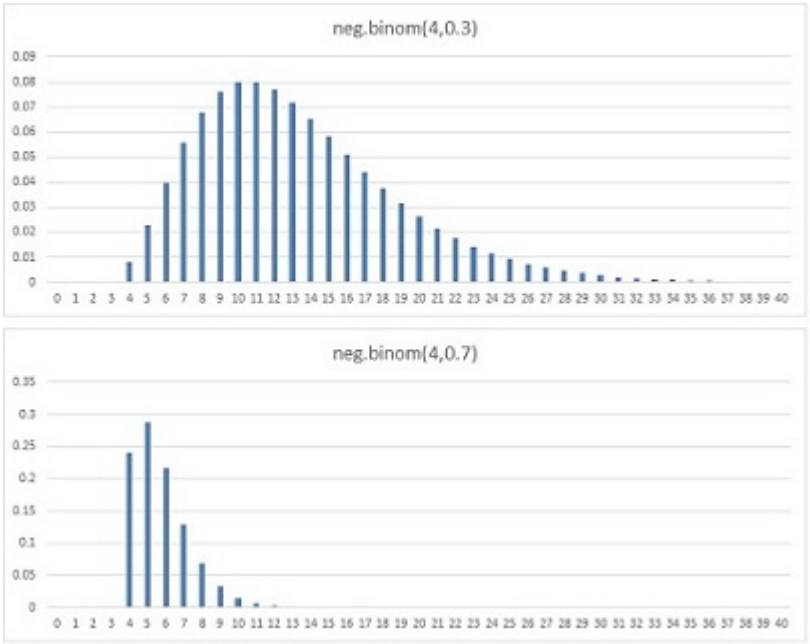


Figure. Top picture: $\mu = 13.3$, $\sigma^2 = 31.1$. Bottom picture: $\mu = 5.7, \sigma^2 = 2.45$.

Remark.
We usually think of a random variable X as making a measurement on the sample point produced by the experiment, and as such, there are typically units associated with this measurement (e.g., $\mu g/\ell$, inches, °F, etc.). By definition, when X is discrete say,

$$\mu = E(X) = \sum_x x\, P(X = x)$$

and, clearly, we see μ is in the same units as X; BUT!

$$\sigma^2 = Var(X) = \sum_x (x - \mu)^2\, P(X = x),$$

and variance σ^2 is in units-squared of the random variable. For this reason, we may also want a measure of spread/fluctuation that is on the same scale as the mean; and we define

$$\sigma = \sqrt{Var(X)}$$

as the ***standard deviation of*** X.

Example. Compute the mean, variance, and standard deviation of X having pmf:

x	-2	-1	0	5
$P(X = x)$	0.4	0.2	0.1	0.3

$$\mu = E(X) = -2(0.4) - 1(0.2) + 0(0.1) + 5(0.3) = 0.5.$$

$$\begin{aligned} \sigma^2 = Var(X) &= E[(X - \mu)^2] \\ &= (-2 - .5)^2(0.4) + (-1 - .5)^2(0.2) + (0 - .5)^2(0.1) + (5 - .5)^2(0.3) \\ &= 2.5 + 0.45 + 0.025 + 6.075 = 9.05. \end{aligned}$$

Advice:
When asked to compute a variance it is usually better to use the following formula:
A computational form for $Var(X)$***:***

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

Remark.
These variance formulas we have given are true for any type of rv; however, the proof below assumes the rv is discrete.

Proof of the computational form of the variance.

$$\begin{aligned}
E[(X-\mu)^2] &= \sum_x (x-\mu)^2\, P(X=x) \\
&= \sum_x (x^2 - 2\mu x + \mu^2)\, P(X=x) \\
&= \sum_x (x^2\, P(X=x) - 2\mu x\, P(X=x) + \mu^2\, P(X=x)) \\
&= \sum_x x^2\, P(X=x) - \sum_x 2\mu x\, P(X=x) + \sum_x \mu^2\, P(X=x) \\
&= \sum_x x^2\, P(X=x) - 2\mu \underbrace{\sum_x x\, P(X=x)}_{=\mu} + \mu^2 \underbrace{\sum_x P(X=x)}_{=1} \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2
\end{aligned}$$

□

Example.
From page 126 we learned if $X \sim$ Bernoulli(p), then $E(X^k) = p$ for all $k \geq 1$. Therefore, in particular, $E(X^2) = p$ and $E(X) = p$, and it follows

$$Var(X) = p - p^2 = p(1-p).$$

Property of Variance.
For any constants a and b,

$$Var(aX+b) = a^2 Var(X).$$

Proof.

$$\begin{aligned}
Var(aX+b) &= E\left[\Big(aX+b-E(aX+b)\Big)^2\right] = E\left[\Big(aX+b-(a\mu+b)\Big)^2\right] \\
&= E\left[a^2(X-\mu)^2\right] = a^2 Var(X).
\end{aligned}$$

Two special cases:

When $a = 1$ and b is any constant, then $Var(X+b) = Var(X)$. This says the variance of X doesn't change if you shift the distribution by a constant b, This seems plausible.

When a is constant and $b = 0$, then $Var(aX) = a^2 Var(X)$. In particular, if $a = -1$, $Var(-X) = Var(X)$.

Example. Let $X \sim \text{binom}(n, p)$. Compute $Var(X)$.

We already know $E(X) = np$. Next,

$$\begin{aligned}
E(X^2) &= \sum_{x=0}^{n} x^2 \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x^2 \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} (x-1+1) \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} (x-1) \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} + \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=2}^{n} (x-1) \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} + np \sum_{x=1}^{n} \underbrace{\frac{(n-1)!}{(x-1)!(n-x)!}}_{=\binom{n-1}{x-1}} p^{x-1} (1-p)^{n-1-(x-1)} \\
&= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!} p^x (1-p)^{n-x} + np \underbrace{\sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{n-1-(x-1)})}_{=1 \text{ by binomial theorem}} \\
&= n(n-1)p^2 \sum_{x=2}^{n} \underbrace{\frac{(n-2)!}{(x-2)!(n-x)!}}_{=\binom{n-2}{x-2}} p^{x-2} (1-p)^{n-2-(x-2)} \quad + \quad np \\
&= n(n-1)p^2 \underbrace{\sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} (1-p)^{n-2-(x-2)}}_{=1 \text{ by binomial theorem}} \quad + \quad np \\
&= n(n-1)p^2 + np.
\end{aligned}$$

Finally,

$$Var(X) = E(X^2) - [E(X)]^2 = n(n-1)p^2 + np - [np]^2 = np(1-p).$$

Summary: If $X \sim \text{binom}(n, p)$, then $\mu = np$, $\sigma^2 = np(1-p)$ and $\sigma = \sqrt{np(1-p)}$. $\square$

Exercise for students.
Use same approach to find $Var(X)$ when $X \sim \text{Poisson}(\lambda)$.

Moment-generating functions.

Suppose X is a random variable with a specified distribution. A very useful function (when it exists) in working with some of the more well-known probability distributions is the so-called moment-generating function:

If the following expected value thought of as a function of θ:

$$M(\theta) = M_X(\theta) = E(e^{\theta X})$$

exists and is finite for θ in an open neighborhood of $\theta = 0$, i.e., $E(e^{\theta X}) < \infty$ for $\theta \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$, then we call $M(\theta)$ the ***moment-generating function (MGF) of*** X or ***of its distribution***.

Some properties of the moment-generating function.

Fact 1. The MGF uniquely identifies the distribution of the random variable.

Fact 2. $|E(X^k)| < \infty$ for all integer $k \geq 1$.

Fact 3. Taking derivatives of the MGF with respect to θ and evaluating at $\theta = 0$ will give the moments of X. For example,

$$M'(0) = E(X), \qquad M''(0) = E(X^2), \qquad M'''(0) = E(X^3), \quad \text{etc.}$$

To roughly see why, since expectation and differentiation are linear operations we can exchange their order:

$$\frac{d^k}{d\theta^k} M(\theta) = \frac{d^k}{d\theta^k} E(e^{\theta X}) = E(\frac{d^k}{d\theta^k} e^{\theta X}) = E(X^k e^{\theta X}) \implies \frac{d^k}{d\theta^k} M(0) = E(X^k).$$

Example. (MGF of a binom(n, p))
Let's compute the MGF (if it exists) of $X \sim \text{binom}(n, p)$. Using LOTUS

$$\begin{aligned} M(\theta) = E(e^{\theta X}) &= \sum_{x=0}^{n} e^{\theta x} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^{\theta})^x (1-p)^{n-x} \\ &= (1 - p + pe^{\theta})^n \qquad \text{(from the binomial theorem.)} \end{aligned}$$

Example.
Let $X \sim \text{Poisson}(\lambda)$. Compute the MGF of X.

$$\begin{aligned} M(\theta) &= E(e^{\theta X}) \\ &= \sum_{x=0}^{\infty} e^{\theta x} P(X = x) \\ &= \sum_{x=0}^{\infty} e^{\theta x} \frac{e^{-\lambda} \lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} e^{\theta x} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\left(e^{\theta} \lambda\right)^x}{x!} \\ &= e^{-\lambda} e^{\lambda e^{\theta}} = e^{\lambda(e^{\theta} - 1)} \end{aligned}$$

Exercise. ***Please do this!***
Use the above MGF to compute μ, σ^2 and σ for the Poisson(λ).

Exercise. ***Please do this!***
Use the binomial MGF we computed on page 131 to re-derive the mean, variance and standard deviation of the binom(n, p).

Cumulative distribution functions.

Let X be ***any*** random variable. The ***cumulative distribution function (CDF)*** or, simply, the ***distribution function*** of X is the function

$$F : \mathbb{R} \to [0,1] \quad \text{defined by} \quad F(x) = P(X \le x).$$

Properties of the CDF. (proofs are on next page...)

Property 1. $F = F(x)$ is a monotone nondecreasing function of x; namely,

$$x < y \implies F(x) \le F(y).$$

Property 2. $F = F(x)$ is always a right-continuous function of x; namely,

$$\text{For all real } x, \quad F(x) = \lim_{h \to 0^+} F(x+h) =: F(x+).$$

Property 3. $F = F(x)$ always possesses left limits; namely,

$$\text{For all real } x, \quad \lim_{h \to 0^+} F(x-h) =: F(x-) \ \text{ exists.}$$

Property 4. $\lim_{x \to -\infty} F(x) = 0.$

Property 5. $\lim_{x \to +\infty} F(x) = 1.$

Remark.
In principle, knowing the CDF of a random variable X is enough to compute probabilities like

$$P(a < X \le b), \quad P(a \le X \le b), \quad P(a < X < b), \quad \text{and} \quad P(a \le X < b).$$

$$P(a < X \le b) = F(b) - F(a).$$

$$P(a \le X \le b) = F(b) - F(a-).$$

$$P(a < X < b) = F(b-) - F(a).$$

$$P(a \le X < b) = F(b-) - F(a-).$$

Proof of Property 1.
Suppose $x < y$. Then the event $(X \le x) \subseteq (X \le y)$, and, by monotonicity of probability, $F(x) = P(X \le x) \le P(X \le y) = F(y)$.

Proof of Property 2.
Since $E_n := \bigcap_{k=1}^{n}(X \le x + \frac{1}{k}) = (X \le x + \frac{1}{n}) \downarrow \bigcap_{k=1}^{\infty}(X \le x + \frac{1}{k}) =^{*} (X \le x)$ as $n \to \infty$, by continuity of probability measure from above,

$$P(E_n) = F(x + \frac{1}{n}) \longrightarrow F(x) = P(X \le x) \quad \text{as } n \to \infty.$$

* to see why $\bigcap_{k=1}^{\infty}(X \le x + \frac{1}{k}) = (X \le x)$ we will show

$$(X \le x) \subseteq \bigcap_{k=1}^{\infty}(X \le x + \frac{1}{k}) \quad \text{and} \quad \bigcap_{k=1}^{\infty}(X \le x + \frac{1}{k}) \subseteq (X \le x).$$

Let $\omega \in (X \le x)$. Then $X(\omega) \le x$. This implies $X(\omega) \le x \le x + \frac{1}{k}$ for every $k \ge 1$. So, $\omega \in (X \le x + \frac{1}{k})$ for every integer $k \ge 1$. Therefore, $\omega \in \bigcap_{k=1}^{\infty}(X \le x + \frac{1}{k})$. This proves the first of the two containments above. We'll show the second containment using complements, namely, we'll show if $\omega \in (X > x)$, then $\omega \in \bigcup_{k=1}^{\infty}(X > x + \frac{1}{k})$. To this end, suppose $\omega \in (X > x)$. Then $X(\omega) > x$. Let $\varepsilon = X(\omega) - x > 0$. Since $\frac{1}{k} \to 0$ as $k \to \infty$, there exists K such that $\frac{1}{k} < \varepsilon$ for all $k \ge K$; i.e., eventually, $\frac{1}{k}$ will be smaller than any fixed positive ε. So, in particular, $X(\omega) - x = \varepsilon > \frac{1}{K}$, which in turn implies $X(\omega) > x + \frac{1}{K}$, i.e., $\omega \in (X > x + \frac{1}{K})$ and, therefore, $\omega \in \bigcup_{k=1}^{\infty}(X > x + \frac{1}{k})$.

Proof of Property 3.
Since $A_n := \bigcup_{k=1}^{n}(X \le x - \frac{1}{k}) = (X - \frac{1}{n}) \uparrow \bigcup_{k=1}^{\infty}(X \le x - \frac{1}{k}) =^{**} (X < x)$ as $n \to \infty$, by continuity of probability measure from below,

$$P(A_n) = F(x - \frac{1}{n}) \longrightarrow F(x-) = P(X < x) \quad \text{as } n \to \infty.$$

** to see why $\bigcup_{k=1}^{\infty}(X \le x - \frac{1}{k}) = (X < x)$ we'll show

$$\bigcup_{k=1}^{\infty}(X \le x - \frac{1}{k}) \subseteq (X < x) \quad \text{and} \quad (X < x) \subseteq \bigcup_{k=1}^{\infty}(X \le x - \frac{1}{k}).$$

The first containment follows because if $\omega \in \bigcup_{k=1}^{\infty}(X \le x - \frac{1}{k})$, then $\omega \in (X \le x - \frac{1}{n})$ for some integer $n \ge 1$, i.e., you can only belong to a union if you belong to at least one of the events in the union. But then $X(\omega) \le x - \frac{1}{n} < x$ implies $\omega \in (X < x)$. The second containment take an aribitrary $\omega \in (X < x)$. Then $X(\omega) = x - \varepsilon$ for some positive ε. But then $X(\omega) = x - \varepsilon \le x - \frac{1}{k}$ for all integers $k > \frac{1}{\varepsilon}$, which shows $\omega \in \bigcup_{k=1}^{\infty}(X \le x - \frac{1}{k})$.

Proof of Property 4.
Since X is a random variable, it's real-valued on Ω, and $\bigcap_{n=1}^{\infty}(X \leq -n) = \varnothing$. So, by continuity of probability measure it follows $P(X \leq -n) \to P(\varnothing) = 0$ as $n \to \infty$.

Proof of Property 5.
Since X is a random variable, $\bigcup_{n=1}^{\infty} P(X \leq n) = \Omega$, and, again, by continuity of probability measure, it follows $P(X \leq n) \to P(\Omega) = 1$ as $n \to \infty$.

Example.
Derive the CDF of the discrete random variable having pmf

x	-1	1	2
$p(x)$	0.6	0.2	0.2

.

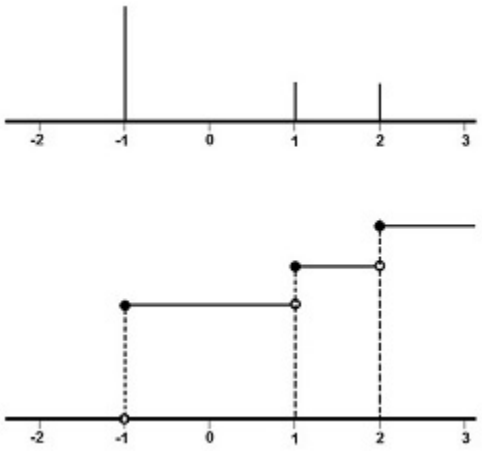


Figure. (Top) pmf (Bottom) CDF

Remark.
Discrete random variables have CDFs that are pure step functions. For any random variable X, $P(X = x) = P(x \leq X \leq x) = F(x) - F(x-)$. So, if X is discrete the pmf of X can be recovered from the CDF of X. The pmf will vanish at all continuity points of the CDF, and, at points of jump discontinuity of F the probability mass $P(X = x) = F(x) - F(x-)$ is just the size of the jump at the point of discontinuity. From the example, the CDF is

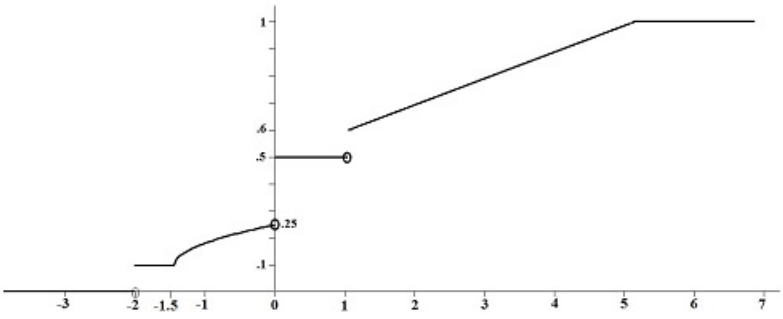
$$F(x) = \begin{cases} 0 & \text{for } x < -1 \\ 0.6 & \text{for } -1 \leq x < 1 \\ 0.8 & \text{for } 1 \leq x < 2 \\ 1 & \text{for } x \geq 2 \end{cases}.$$

Example.
The following is a CDF for a random variable, say, X:

$$F(x) = \begin{cases} 0 & \text{for } x < -2 \\ .1 & \text{for } -2 \le x < -1.5 \\ .1\left(1 + \sqrt{2.25 - x^2}\right) & \text{for } -1.5 \le x < 0 \\ .5 & \text{for } 0 \le x < 1 \\ .6 + .1(x-1) & \text{for } 1 \le x < 5 \\ 1 & \text{for } x \ge 5. \end{cases}$$

Graph this CDF and then use it to compute each of the following:
(a) $P(X = 1)$
(b) $P(X = -1)$
(c) $P(0 < X < 3)$
(d) $P(0 \le X \le 2)$

SOLUTION:



(a) $P(X = 1) = F(1) - F(1-) = .6 + .1(1-1) - .5 = \mathbf{.1}$.

(b) $P(X = -1) = F(-1) - F(-1-) = \mathbf{0}$ since F is continuous at $x = -1$.

(c) $P(0 < X < 3) = F(3-) - F(0) = F(3) - F(0) = .6 + .1(3-1) - .5 = \mathbf{.3}$.

(d) $P(0 \le X \le 2) = F(2) - F(0-) = .6 + .1(2-1) - .1\left(1 + \sqrt{2.25 - (0)^2}\right) = .7 - .25 = \mathbf{.45}$.

Remark.
The random variable X having the CDF in this last example cannot be discrete since the CDF is not a step function. We'll see shortly that X cannot be continuous either; therefore, this rv X is neither discrete nor continuous. The important point here is that, with this rv, we computed probabilities through its CDF.

IV. Continuous random variables.

Continuous random variables.
Discrete random variables have CDFs that are pure steps functions. Continuous random variables have CDFs that are ***continuous functions***.

A random variable X will be called a ***continuous random variable*** if its CDF is a continuous function on the real line.

Let's assume we have a continuous random variable X and let

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

be its continuous CDF. Fix an arbitrary $x \in \mathbb{R}$. Then

$$P(X = x) = P(x \leq X \leq x) = F_X(x) - F_X(x-) = 0$$

since $F(x)$ being continuous means the left limit $F(x-)$ at x equals the value of the function $F(x)$ ar x. And we arrive at a strange feature of continuous random variables. They assign ***zero*** probability mass to individual points.

Now, although

$$P(x - h < X \leq x) = F_X(x) - F_X(x - h) \longrightarrow 0 \quad \text{as } h \downarrow 0, \tag{5}$$

equation (5) gives us some hope; namely, if the CDF is continuously differentiable (more than just continuous), then

$$\frac{P(x - h < X \leq x)}{h} = \frac{F_X(x) - F_X(x - h)}{h} \longrightarrow f(x) \quad \text{as } h \downarrow 0.$$

Of course, the function $f(x) =: F'(x)$ and is measuring the probability mass per unit value at x, namely, $f(x)$ is the ***probability density function (pdf)*** at x.

Assumption about continuous rvs we will make:
Throughout this course ***we will assume*** the CDF admits the existence of a pdf $f(x)$.

What can we do with the pdf when we have it ?

Let $-\infty < a < b < \infty$ and take $n > 1$ to be large. Set $h = \frac{b-a}{n}$ so that

$$(a, b] = (a, a + h] \cup (a + h, a + 2h] \cup \cdots \cup (a + (n - 1)h, b] .$$

In what follows I am assuming the pdf $f(x)$ is continuous but this is not needed.

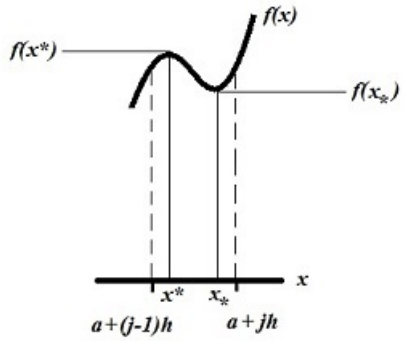


Figure. On $\Big(a+(j-1)h, a+jh\Big]$, minimizer $= x_*$, maximizer $= x^*$

$$\begin{aligned} P(a < X \le b) &= \sum_{j=1}^{n} P\left(a+(j-1)h < X \le a+jh\right) \\ &= \sum_{j=1}^{n} F\left(a+jh\right) - F\left(a+(j-1)h\right) \\ &= \sum_{j=1}^{n} \frac{F\left(a+jh\right) - F\left(a+(j-1)h\right)}{h} \cdot h \\ &= \sum_{j=1}^{n} \underbrace{\frac{F\left(a+jh\right) - F\left(a+jh-h\right)}{h}}_{\approx f(x)} \cdot h \end{aligned}$$

Therefore,

$$\sum_{j=1}^{n} f(x_*) \cdot h \le P(a < X \le b) \le \sum_{j=1}^{n} f(x^*) \cdot h,$$

and, as $n \to \infty$, we get the Riemann integral of $f(x)$ over the interval $(a, b]$:

$$P(a < X \le b) = \int_a^b f(x)\, dx.$$

Properties of a pdf.

1. $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\,du$ - this is the first fundamental theorem of calculus.
2. $f(x) = F'(x)$ - this is the second fundamental theorem of calculus.
3. Since the CDF is defined on the entire real line, so is the pdf.
4. Since the CDF is monotone nondecreasing, $f(x) \geq 0$ for all $x \in \mathbb{R}$.
5. Since $\lim_{x\to\infty} F(x) = 1$, $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

Remark.
Properties $3, 4$ and 5 characterize a pdf; i.e., if a function $f : \mathbb{R} \to [0, \infty)$ is such that $\int_{-\infty}^{\infty} f(x)\,dx = 1$, then there is a random variable having $f(x)$ as its pdf.

Remark. (computing probabilities of a continuous rv with given pdf)
For any $a \leq b$, $P(a < X \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx = \int_{a}^{b} f(x)\,dx$, and this allows us to compute probabilities when the pdf $f(x)$ is known. Moreover, because the pdf puts *no* probability mass on individual values of the rv, $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$.

Advice:
It's helpful to remember, like in a course in Physics, that when computing probabilities involving discrete random variables we add probability masses to get probability mass; when computing probabilities involving continuous random we integrate probability density to get probability mass.

Example.
A researcher believes the pdf of a continuous rv X has the shape of the following function:

$$f(x) = \begin{cases} cx(1-x) & \text{for } 0 \le x \le 1 \\ 0 & \text{elsewhere} \end{cases}.$$

(a) Find the constant c that makes this a pdf.
(b) Compute $P(\frac{1}{3} \le X \le \frac{2}{3})$.

SOLUTION:
(a) $f(x) = 0$ for all $x \notin [0,1]$, and, for $x \in [0,1]$ and any fixed constant c, $cx(1-x)$ doesn't change sign. So $f(x)$ will be ≥ 0 for all real x when $c \ge 0$. We now look for c that makes the total integral 1:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,dx &= \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx \\ &= \int_{-\infty}^{0} 0\,dx + \int_{0}^{1} cx(1-x)\,dx + \int_{1}^{\infty} 0\,dx \\ &= c\int_{0}^{1} x - x^2\,dx = c\Big(\frac{x^2}{2} - \frac{x^3}{3}\Big|_{x=0}^{x=1}\Big) = \frac{c}{6} = 1 \implies c = 6. \end{aligned}$$

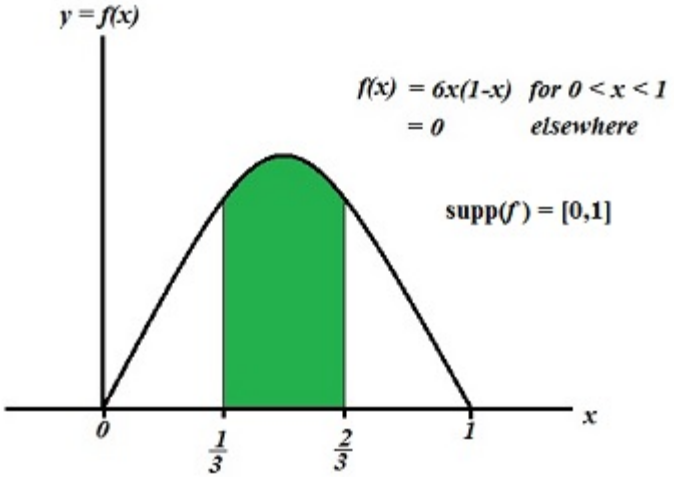


Figure. Graph of the pdf in this example. Green shaded area is $P(\frac{1}{3} \le X \le \frac{2}{3})$.

(b) $P(\frac{1}{3} \le X \le \frac{2}{3}) = \int_{\frac{1}{3}}^{\frac{2}{3}} 6x - 6x^2\,dx = 3x^2 - 2x^3\big|_{x=\frac{1}{3}}^{x=\frac{2}{3}} = \frac{4}{3} - \frac{16}{27} = \frac{20}{27}$.

Remark.
Many of the pmfs for the named discrete distributions were modeled from discrete experiments: Bernoulli trials, sampling with or without replacement, or repeated trials of such. As for continuous random variables, we will work with several *very useful* pdfs that have been modeled to mimic many of these discrete pmfs and/or their properties to the continuous realm. Here's a list of some notable ones we will discuss:

uniform(a, b) - continuous analog to the discrete uniform to the interval $[a, b]$
exponential(λ) - continuous analog to the geometric distribution
Gamma(α, β) - continuous analog to the negative binomial distribution
χ^2_n (chi-square distribution with n degrees of freedom) - just a Gamma$(\frac{n}{2}, 2)$
Normal(μ, σ^2) distribution
Pareto distribution - continuous analog to the Zeta distribution
Laplace/double exponential
Beta(α, β)
t- and F-distributions
bivariate normal distribution
Dirichlet distribution
many others as time permits.

Example. (the ***exponential***(λ) ***distribution***)
Suppose X is a continuous random variable having pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{elsewhere} \end{cases}.$$

We will write this as $X \sim \exp(\lambda)$. This distribution is a widely used model of component lifetimes, X, i.e., where X measures the how long a component lives. If $X > x$ then this means the component is still "alive" at time x. The parameter $\lambda > 0$ is the reciprocal of the *mean lifetime* of the component. As a concrete example, suppose the component is an incandescent lightbulb having a quoted mean lifetime of 10 (thousand hours) so that $\lambda = \frac{1}{10}$. In this case, $f(x) = \frac{1}{10} e^{-\frac{x}{10}}$ for $x \geq 0$.

If $X \sim \exp(\frac{1}{10})$, compute (a) $P(X > 10)$, (b) $P(5 < X < 15)$, and (c) $P(-5 < X < 10)$.

SOLUTION:
(a) This problem is asking us to compute the probability the component lasts longer that its mean lifetime.

$$P(X > 10) = \int_{10}^{\infty} \frac{1}{10} e^{-\frac{x}{10}}\, dx = -e^{-\frac{x}{10}}\Big|_{x=10}^{\infty} = 0 - (-e^{-\frac{10}{10}}) = e^{-1} \approx .3679\ldots.$$

This says about 63.21% of the components (the majority) will not last 10,000 hours.

(b) $P(5 < X < 15) = \int_5^{15} \frac{1}{10} e^{-\frac{x}{10}}\, dx = -e^{-\frac{x}{10}}\Big|_{x=5}^{x=15} = e^{-\frac{1}{2}} - e^{-\frac{3}{2}} \approx 0.3834.$

(c) $P(-5 < X < 10) = \int_{-5}^{10} f(x)\, dx = \int_{-5}^{0} 0\, dx + \int_0^{10} \frac{1}{10} e^{-\frac{x}{10}}\, dx = P(X \leq 10) = .6321\ldots$ from part (a). Just notice that $(-5, 10)$ crosses over into the part of the domain where the density vanishes.

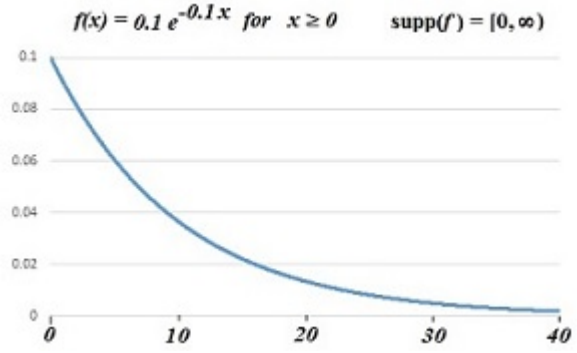


Figure. Graph of the $\exp(\frac{1}{10})$ pdf of the example.

Indicator notation for piecewise functions.
Probability density functions more often than not are piecewise defined, i.e., the functions change their rules over different portions of the real line. The $\exp(\lambda)$ pdf is one example: the pdf equals $\lambda e^{-\lambda x}$ for $x \geq 0$, but $= 0$ for $x < 0$. On the previous page I decided to write this pdf like this:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} .$$

But, there is a rather clever alternate way to write this pdf using ***indicator function notation***, which I now describe.

Let A be a set. Define the ***indicator function*** as

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

With this notation, we can write our $\exp(\lambda)$ pdf *succinctly* like this:

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{[0,\infty)}(x).$$

The presence of the indicator function just tells us that if $x \in [0, \infty)$, it multiplies by 1 and we get $\lambda e^{-\lambda x}$; and, if $x \notin [0, \infty)$, it multiplies by 0 and we just get the *zero* function.

We have a teaching assistant who fancies using this notation, so you may see it used often in certain circles, but I tend to use the notation on the previous page usually spelling out the piecewise definitions and just keeping track of the domains of definition especially in the univariate case (i.e., just one rv case). Nevertheless, there is a benefit to using this notation, especially when we start working with joint pdfs (more than one rv), and we may revisit the use of this notation later in the course.

Exercise for the student.
Let $X \sim \exp(\lambda)$.
(a) Show that for *any* $t > 0$, $P(X > t) = e^{-\lambda t}$.
(b) Show the exponential distribution has this peculiar property:
the ***memoryless property***:

$$\text{For any } s, t > 0, P(X > s + t | X > s) = P(X > t).$$

In words this property says: the probability a component survives an additional t units of time given it has survived s units of time *is the same as* the probability a "new" component survives t unit of time, i.e., *the component has no memory of its age!*

Expected values of continuous random variables.
If X is a continuous random variable with pdf $f(x)$, we define the ***expected value*** to be

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

when it exists. Recall the discussion on the existence of expected values on page 122: $E(X)$ exists when $E(|X|) = \int_{-\infty}^{\infty} |x| f(x)\, dx$ is finite.

Advice:
The integration in this expected value need only be done over the support of the pdf because the pdf (and, hence, the integrand) will vanish for values not in the support.

Remark.
A classic example of a continuous random variable whose expected value doesn't exist is a rv having the **(standard) *Cauchy pdf***:

$$f(x) = \frac{1}{\pi(1+x^2)} \quad \text{for } -\infty < x < \infty.$$

In this case

$$E(X^+) = E(X^-) = \frac{1}{\pi}\int_0^{\infty} x \cdot \frac{1}{1+x^2}\, dx = \frac{1}{2\pi}\left(\ln(1+x^2)\Big|_{x=0}^{\infty}\right) = +\infty.$$

Example.
Suppose $X \sim \exp(\lambda)$. Compute $E(X)$.

SOLUTION:
Recall the pdf of the $\exp(\lambda)$ is $f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$. The support of this pdf is $[0, \infty)$. Therefore,

$$\begin{aligned}
\mu = E(X) &= \int_{-\infty}^{\infty} x f(x)\, dx \\
&= \int_{-\infty}^{0} x \cdot 0\, dx + \int_{0}^{\infty} x \cdot \lambda e^{-\lambda x}\, dx \quad \text{(only need integrate over support of } f\text{)} \\
&= \lambda \int_{0}^{\infty} \underbrace{x}_{=:u} \underbrace{e^{-\lambda x}\, dx}_{=:dv} \qquad \text{(use an integration by parts)} \\
&= \underbrace{-x e^{-\lambda x}\Big|_{x=0}^{\infty}}_{=0} - \int_{0}^{\infty} -e^{-\lambda x}\, dx \\
&= \int_{0}^{\infty} e^{-\lambda x}\, dx \\
&= -\frac{e^{-\lambda x}}{\lambda}\Big|_{x=0}^{\infty} = \frac{1}{\lambda}.
\end{aligned}$$

We've showed that the ***mean of an exp***(λ) ***is the reciprocal of*** λ: $\mu = \frac{1}{\lambda}$.

The law of the unconscious statistician that we learned for discrete rvs also holds true for continuous rvs. A proof, however, is a bit beyond the scope of this course.

Law of the Unconscious Statistician (LOTUS) - continuous case
Suppose X is a continuous random variable with pdf $f(x)$ and $g = g(x)$ is a function. Then

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

assuming this expected value exists.

Example. (the ***uniform***(a,b) ***distribution***)
Let $-\infty < a < b < \infty$. A continuous rv X is said to be ***uniformly distributed on the interval*** $[a,b]$ if the pdf of X is given by

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{elsewhere} \end{cases},$$

or, using the *indicator notation*: $f(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x)$. This distribution is the continuous analog of equally likely: drawing an observation from the interval $[a,b]$ equally likely at random. Please compute the mean and variance of the uniform(a,b).

SOLUTION:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\,dx = \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{x^2}{2}\Big|_{x=a}^{x=b} \\ &= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}, \end{aligned}$$

and using LOTUS,

$$\begin{aligned} E(X^2) &= \int_a^b x^2 \frac{1}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{x^3}{3}\Big|_{x=a}^{x=b} \\ &= \frac{b^3-a^3}{3(b-a)} = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3}. \end{aligned}$$

Finally,

$$Var(X) = E(X^2) - \big(E(X)\big)^2 = \frac{b^2+ab+a^2}{3} - \big(\frac{a+b}{2}\big)^2 = \cdots = \frac{(b-a)^2}{12},$$

the basic algebra involved is left to the reader.

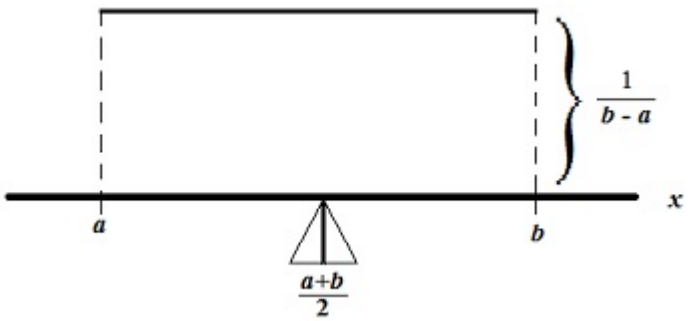


Figure. Graph of the pdf of a uniform(a,b). Mean is the midpoint of the interval $[a,b]$, $\mu = \frac{a+b}{2}$, which should not be surprising – it's exactly where you'd think the fulcrum of the see-saw should go!

Moment-generating functions - revisited
For some pdfs, the moment generating function will exist, and when it does it can be very useful.

Recall the definition from page 131:
If the following expected value thought of as a function of θ:

$$M(\theta) = M_X(\theta) = E(e^{\theta X})$$

exists and is finite for θ in an open neighborhood of $\theta = 0$, i.e., $E(e^{\theta X}) < \infty$ for $\theta \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$, then we call $M(\theta)$ the ***moment-generating function (MGF) of*** X or ***of its distribution***.

Example.
Compute the moment generating function of $X \sim \exp(\lambda)$.

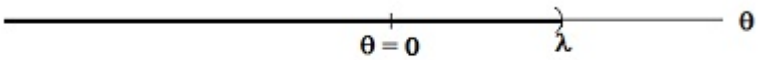
SOLUTION:

$$\begin{aligned} M(\theta) &= E\left(e^{\theta X}\right) \\ &= \int_0^\infty e^{\theta x} \cdot \lambda e^{-\lambda x}\, dx \\ &= \lambda \int_0^\infty e^{-(\lambda-\theta)x}\, dx \qquad \text{(important to have } \theta < \lambda \text{ in this step!)} \\ &= -\frac{\lambda}{\lambda-\theta} e^{-(\lambda-\theta)x}\Big|_{x=0}^{\infty} \\ &= \frac{\lambda}{\lambda-\theta}, \end{aligned}$$

and, sometimes, it is convenient to write this as

$$M(\theta) = \left(1 - \frac{\theta}{\lambda}\right)^{-1}.$$

It is important to reiterate that this MGF is only defined for values of $\theta < \lambda$, but since $\lambda > 0$ this still provides us an open neighborhood of $\theta = 0$ where $M(\theta)$ exists and is finite. So this is the MGF of the $\exp(\lambda)$ distribution.



MGF of the exp(λ) exists and is finite in a neighborhood of θ = 0.

Remark. (The MacLaurin expansion of the MGF)
Recall the MacLaurin expansion

$$e^{\theta X} = 1 + \theta X + \frac{\left(\theta X\right)^2}{2!} + \frac{\left(\theta X\right)^3}{3!} + \frac{\left(\theta X\right)^4}{4!} + \cdots .$$

Taking the expected value on both sides and using linearity of expectation, it follows

$$M(\theta) = 1 + E(X)\theta + E(X^2)\frac{\theta^2}{2!} + E(X^3)\frac{\theta^3}{3!} + E(X^4)\frac{\theta^4}{4!} + \cdots ,$$

and we see the kth moment $E(X^k)$ is really just the coefficient of the $\frac{\theta^k}{k!}$-term in this MacLaurin expansion.

Remark.(General formula for the moments of an $\exp(\lambda)$)
From the geometric series formula, if $\left|\frac{\theta}{\lambda}\right| < 1$, then the MGF of the $\exp(\lambda)$ can be written as

$$\begin{aligned} M(\theta) &= \left(1 - \frac{\theta}{\lambda}\right)^{-1} = 1 + \frac{\theta}{\lambda} + \left(\frac{\theta}{\lambda}\right)^2 + \left(\frac{\theta}{\lambda}\right)^3 + \left(\frac{\theta}{\lambda}\right)^4 + \cdots \\ &= 1 + \frac{1}{\lambda}\theta + \frac{2!}{\lambda^2}\frac{\theta^2}{2!} + \frac{3!}{\lambda^3}\frac{\theta^3}{3!} + \frac{4!}{\lambda^4}\frac{\theta^4}{4!} + \cdots . \end{aligned}$$

On the other hand, from the previous remark

$$M(\theta) = 1 + E(X)\theta + E(X^2)\frac{\theta^2}{2!} + E(X^3)\frac{\theta^3}{3!} + E(X^4)\frac{\theta^4}{4!} + \cdots .$$

So, by matching coefficients of like-powers of θ, we see when $X \sim \exp(\lambda)$,

$$E(X^k) = \frac{k!}{\lambda^k}.$$

In particular, we immediately get

$$E(X) = \frac{1}{\lambda}, \quad E(X^2) = \frac{2}{\lambda^2} \quad \Longrightarrow \quad Var(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

Other moment-related quantities.

The z-score.
Suppose X is a random variable with mean μ and standard deviation σ.
We define the z-score of X to be

$$Z := \frac{X - \mu}{\sigma},$$

and it measures the number of standard deviations X is away from its mean.
Note that the numerator and denominator are both in units of the random variable and, therefore, the z-score is a dimensionless quantity - useful for comparing different random variables and distributions.

Centered moments.
If X is a random variable with mean μ, then $E\big((X-\mu)^k\big)$ is called the kth centered (or central) moment. The variance $\sigma^2 = Var(X)$ is the 2nd centered moment.

Remark.
I've mentioned that the moments of a probability distribution reveal information about the distribution - the more moments we know, the more information we have about the distribution. The 1st moment is the mean and tells us where the distribution is *located* on the real line, the 2nd centered moment is the variance and tells us how spread out the values of the random variable are from its mean. There are two other measures related to higher moments that I now want to tell you about. They are the ***skewness*** of the distribution and the ***kurtosis*** of the distribution.

Skewness and Kurtosis.
Let X be a random variable with mean μ and standard deviation σ. We define the ***skewness*** of the distribution of X by

$$\tilde{\mu}_3 := E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right),$$

and is a measure of asymmetry in the distribution; we define the ***kurtosis**** of the distribution of X by

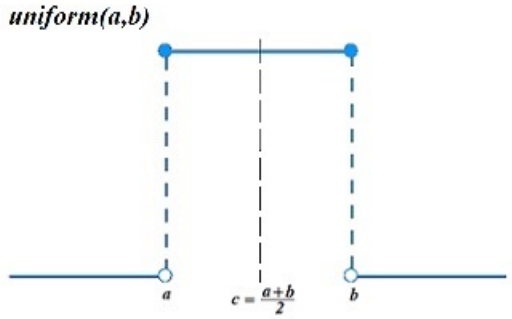
$$\tilde{\mu}_4 := E\left(\left(\frac{X-\mu}{\sigma}\right)^4\right),$$

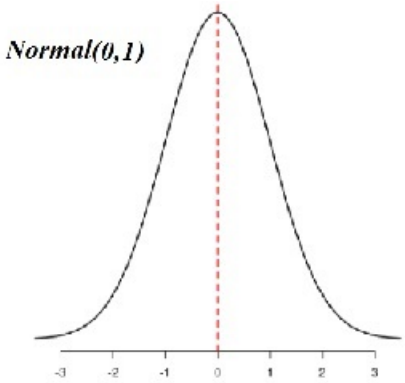
and is a measure of how "heavy" the "tails" of the distribution are.
* The kurtosis is usually a value which is compared to that of the normal distribution which has kurtosis equal to 3. In fact, some authors define the kurtosis instead by $\tilde{\mu}_4 - 3$, i.e., subtracting 3 from the quantity given, which, when positive, tells us the tails are heavier than that of a normal distribution, and, when negative, that tails are lighter than that of a normal distribution.

Remark. (skewness vs. actual symmetry)
The concept of symmetry and the measure we are calling skewness are *different*. A continuous random variable X with pdf f is said to be ***symmetric*** if the graph of the pdf is symmetric about some line $x = c$, i.e., if there is a constant c such that

$$f(c - x) = f(c + x) \quad \text{for all } x \geq 0.$$

If the pdf has a mean μ that exists and is finite, then the value of c above is μ. However, note that the Cauchy pdf on page 145 is symmetric because $f(0 + x) = f(0 - x)$ for all $x \geq 0$, but the Cauchy has no mean and, therefore, $\tilde{\mu}_3$ is not defined for the Cauchy! Here are pictures of some symmetric pdfs:





It is a fact that if a pdf is symmetric and possesses a third moment, then the skewness $\tilde{\mu}_3 = 0$. Can you prove this? However, there are pdfs with $\tilde{\mu}_3 = 0$ that are *not* symmetric.

Exercise for the student.
For the exp(λ) pdf, verify $\tilde{\mu}_3 = 2$ and $\tilde{\mu}_4 = 9$.
The skewness $\tilde{\mu}_3 = 2 > 0$ means the exp(λ) distribution is *positively skewed* which loosely means the random variable is more likely to take on larger values than smaller values (see the exp(0.1) pdf in the figure on page 143). The kurtosis $\tilde{\mu}_4 = 9$ is greater than 3 which means the exp(λ) distribution has heavier tails than the normal distribution.

In order to introduce commonly used pdfs in probability, we need to take this short mathematical digression and introduce Euler's Gamma function.

Euler's Gamma function.
For real $\alpha > 0$, we define the function

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u}\, du.$$

This function is, in fact, defined for values of all α in the complex plane (with the exception of the zero and negative even integers), but, in this course, we only need the behavior of this function for positive real values α.

Let's compute $\Gamma(\alpha)$ for some values of $\alpha > 0$.

$\alpha = 1$:

$$\Gamma(1) \;=\; \int_0^\infty u^{1-1} e^{-u}\, du = \int_0^\infty e^{-u}\, du = 1.$$

$\alpha = 2$:

$$\begin{aligned}
\Gamma(2) &= \int_0^\infty u^{2-1} e^{-u}\, du \\
&= \int_0^\infty u e^{-u}\, du \\
&= \underbrace{-ue^{-u}\big|_{u=0}^{\infty}}_{=0} + \underbrace{\int_0^\infty e^{-u}\, du}_{=:\Gamma(1)=1} \qquad \text{(used an integration by parts)} \\
&= 1.
\end{aligned}$$

In fact, when $\alpha > 0$:

$$\begin{aligned}
\Gamma(\alpha+1) &= \int_0^\infty u^{\alpha} e^{-u}\, du \\
&= \underbrace{-u^{\alpha}e^{-u}\big|_{u=0}^{\infty}}_{=0} + \alpha \underbrace{\int_0^\infty u^{\alpha-1} e^{-u}\, du}_{=:\Gamma(\alpha)} \qquad \text{(used an integration by parts)} \\
&= \alpha\Gamma(\alpha),
\end{aligned}$$

and we arrive at a very useful relationship...

The Reduction formula of the Gamma function.
For $\alpha > 0$,

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha).$$

Remark. (Euler's Gamma generalizes the notion of *factorial* to positive reals)
The reduction formula can also be stated as $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, but we would now require $\alpha > 1$ since $\Gamma(\alpha - 1)$ only makes sense when $\alpha - 1 > 0$. The reduction formula is very useful because it allows us to compute $\Gamma(x)$ for any positive real x as a function on $\Gamma(y)$, where $y \in (0, 1]$:

$$\begin{aligned}\Gamma(5) &= 4\Gamma(4) = 4 \cdot 3\Gamma(3) = 4 \cdot 3 \cdot 2\Gamma(2) = 4 \cdot 3 \cdot 2 \cdot 1 \underbrace{\Gamma(1)}_{=1}. \\ &= 4!\end{aligned}$$

and, for integer $n \geq 1$,

$$\Gamma(n) = (n-1)!.$$

In this sense, we see Euler's Gamma function agrees with the factorial on integers. Even more:

$$\Gamma(4.6) = (3.6)(2.6)(1.6)(0.6)\Gamma(0.6),$$

and, if needed, we can look up the value of $\Gamma(0.6)$ in a table.[3]

Remark.
We will show a bit later that

$$\Gamma(\frac{1}{2}) = \sqrt{\pi}.$$

When we combine this result with the last remark, for instance,

$$\Gamma(\frac{7}{2}) = \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2}\Gamma(\frac{1}{2}) = \frac{15\sqrt{\pi}}{8}.$$

End of digression.

Remark. (creating pdfs from nonnegative functions with finite integral)
If we have any real-valued function $g = g(x)$ defined on the real line which is nonnegative and whose integral is finite and positive, say,

$$\int_{-\infty}^{\infty} g(x)\, dx = c > 0,$$

then, in fact $f(x) := \frac{g(x)}{c}$ will be a pdf – just divide both sides of the display by c. Many pdfs are created this way: find a nonnegative function whose graph/shape mimics how you want a pdf to behave, then just normalize it to have a total integral equal to 1.

[3]Abramowitz & Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Dover (1965).*

Following the last remark, let's consider the function

$$\begin{aligned} g(x) &= x^{\alpha-1}e^{-x/\beta} \quad \text{for } x > 0 \\ &= 0 \quad \text{for } x \leq 0, \end{aligned}$$

where $\alpha > 0$ and $\beta > 0$ are fixed constants.

The support of this function is $[0, \infty)$ and its integral is

$$\begin{aligned} \int_0^\infty g(x)\,dx &= \int_0^\infty x^{\alpha-1}e^{-x/\beta}\,dx \qquad \text{(make substitution } u = x/\beta) \\ &= \int_0^\infty (\beta u)^{\alpha-1}e^{-u}\,\beta\,du \\ &= \beta^\alpha \int_0^\infty u^{\alpha-1}e^{-u}\,du \\ &= \beta^\alpha \Gamma(\alpha). \end{aligned}$$

Therefore, we have

The Gamma(α, β) pdf.
The function

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} & \text{for } x > 0 \\ \\ 0 & \text{for } x \leq 0 \end{cases}$$

is a pdf. In this context, the parameter α is called the ***shape*** parameter and β is called the ***scale*** parameter.

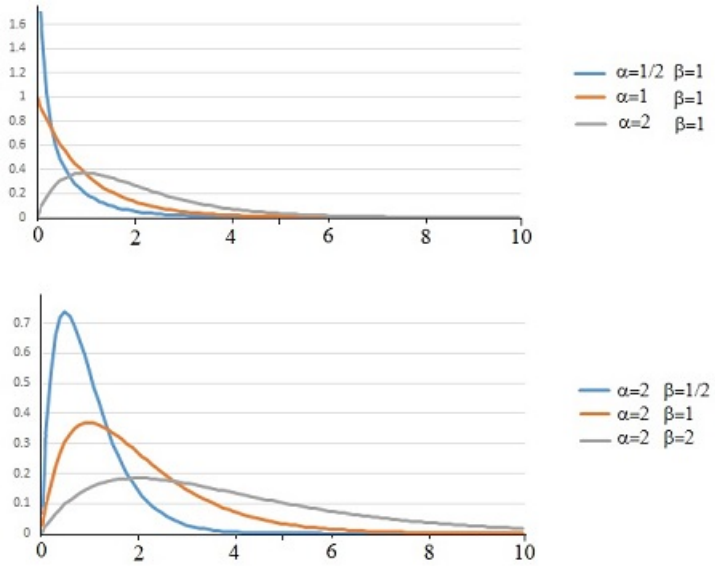


Figure. Top picture: scale fixed at $\beta = 1$ and shape α changes from $\frac{1}{2}$ to 1 to 2. Bottom picture: shape fixed at $\alpha = 2$ and scale β changes from $\frac{1}{2}$ to 1 to 2.

As the figure illustrates, if we fix the scale parameter, then changing the shape parameter changes the essential shape of the graph: when $0 < \alpha < 1$ the graph has a vertical asymptote at $x = 0$; when $\alpha = 1$, the graph is an $\exp(\frac{1}{\beta})$ pdf taking the value $\frac{1}{\beta}$ at $x = 0$; when $\alpha > 1$, the functions take the value 0 at $x = 0$.

If we fix the shape parameter, then, as the scale parameter increases (decreases), the graph of the function gets stretched out (scrunched in). I.e., β genuinely acts as a scale parameter should: when it increases, we increase the scale; when it decreases, we decrease the scale.

Remark. (Inconsistencies among various definitions of the Gamma distribution)
Some authors prefer to use a "rate" parameter λ instead of the scale parameter β when defining their Gamma distribution. These authors define their Gamma(α, λ) pdf like this:

$$f(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases},$$

and, in this form, when $\alpha \in \mathbb{Z}_+$, it's also called the ***Erlang distribution***.

But notice when $\alpha = 1$, the Gamma$(1, \lambda)$ reduces to the pdf $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, i.e., the Gamma$(1, \lambda) \equiv \exp(\lambda)$; whereas, by *our* definition, a Gamma$(1, \frac{1}{\lambda}) \equiv \exp(\lambda)$. So take note: what these authors are calling a Gamma(α, λ) is really what we are calling a Gamma$(\alpha, \frac{1}{\lambda})$, and, what we're calling a Gamma(α, β), these other authors would be calling a Gamma$(\alpha, \frac{1}{\beta})$. Know whether or not you are dealing with a person who uses a rate parameter instead of a scale parameter in the definition of their Gamma!

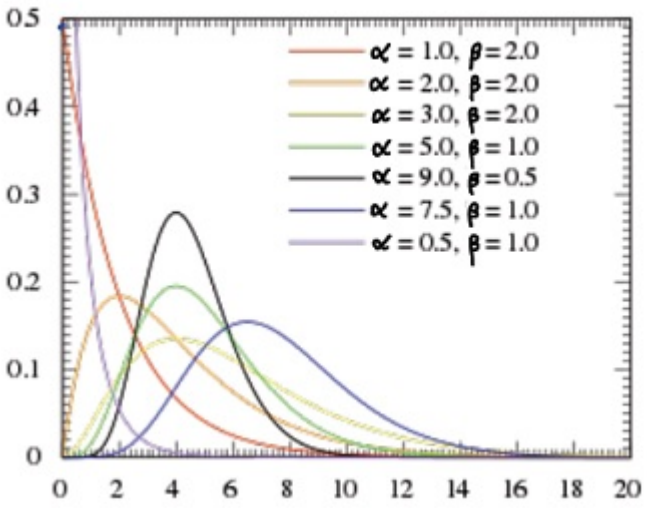


Figure. Some more graphs of Gamma(α, β) densities for various choices of α and β.

Surprisingly, we will not be interested much in finding areas under these curves; although, it can be done by numerical integration methods such as trapezoidal rule or Simpson's rule if needed. Instead, in this course, we'll be much more interested in the *global behavior* of the Gamma distribution - the mean, variance, moments, and MGF.

Normalization "tricks".

Before we investigate the global behavior of the Gamma(α, β) distribution, I wish to introduce a ***normalization trick*** which will allow us to *quickly* compute integrals by recognizing we are integrating the functional form of a known pdf over its entire support. Although I will introduce the idea with the Gamma(α, β) pdf, the idea extends to virtually any pdf (and pmf) we work with in this course.

Here's the idea for the Gamma(α, β) distribution. For $\alpha > 0$ and $\beta > 0$, we know

$$\int_0^\infty \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}\, dx = 1,$$

because pdfs integrate to 1 over their entire support. $\beta^\alpha\Gamma(\alpha)$ is the normalizing constant, so, when multiplied to the other side, we obtain the ***extremely useful fact:***

$$\int_0^\infty x^{\alpha-1}e^{-x/\beta}\, dx = \beta^\alpha\Gamma(\alpha).$$

In words: when we integrate the functional form of the Gamma(α, β) density, i.e., $x^{\alpha-1}e^{-x/\beta}$, over its entire support $(0, \infty)$ we get the normalizing constant, $\beta^\alpha\Gamma(\alpha)$: *mnemonic*: "scale to the shape times Gamma of shape".

Examples.
Compute each of the following integrals:

(a) $\displaystyle\int_0^\infty x^3 e^{-x/2}\, dx$

(b) $\displaystyle\int_0^\infty x^{\frac{1}{2}} e^{-2x}\, dx$

(c) For $n \in \mathbb{Z}_+$, simplify $\displaystyle\int_0^\infty x^n e^{-nx}\, dx$.

SOLUTIONS:

(a) $\displaystyle\int_0^\infty \underbrace{x^3 e^{-x/2}}_{\alpha=4,\ \beta=2}\, dx = 2^4\Gamma(4) = 16(3!) = 96.$

(b) $\displaystyle\int_0^\infty \underbrace{x^{\frac{1}{2}} e^{-2x}}_{\alpha=\frac{3}{2},\ \beta=\frac{1}{2}}\, dx = \left(\frac{1}{2}\right)^{\frac{3}{2}}\Gamma(\frac{3}{2}) = \left(\frac{1}{2}\right)^{\frac{3}{2}}\frac{1}{2}\underbrace{\Gamma(\frac{1}{2})}_{\sqrt{\pi}} = \left(\frac{1}{2}\right)^{\frac{5}{2}}\sqrt{\pi} = \frac{1}{4}\sqrt{\frac{\pi}{2}}.$

(c) I get $\frac{(n-1)!}{n^n}$... check my work!

Exercise for the student.

Compute $\displaystyle\int_0^\infty \frac{e^{-x/2}}{\sqrt{x}}\, dx.$

Moments of the Gamma(α, β).

With the normalization trick and the basic properties of Euler's Gamma function we are now ready to compute the moments of the Gamma(α, β). We'll only compute the first two moments of this distribution for now.

The mean and variance of $X \sim$ Gamma(α, β):

$$\begin{aligned} E(X) &= \int_0^\infty x \cdot \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}\, dx \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty \underbrace{x^\alpha e^{-x/\beta}}_{\substack{\text{shape}=\alpha+1,\\ \text{scale}=\beta}}\, dx \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \cdot \beta^{\alpha+1}\Gamma(\alpha+1) = \frac{\beta^{\alpha+1}\alpha\Gamma(\alpha)}{\beta^\alpha\Gamma(\alpha)} = \alpha\beta. \end{aligned}$$

$$\begin{aligned} E(X^2) &= \int_0^\infty x^2 \cdot \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}\, dx \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty \underbrace{x^{\alpha+1} e^{-x/\beta}}_{\substack{\text{shape}=\alpha+2,\\ \text{scale}=\beta}}\, dx \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \cdot \beta^{\alpha+2}\Gamma(\alpha+2) = \frac{\beta^{\alpha+2}(\alpha+1)\alpha\Gamma(\alpha)}{\beta^\alpha\Gamma(\alpha)} = \alpha(\alpha+1)\beta^2. \end{aligned}$$

From here, the variance is immediate:

$$Var(X) = E(X^2) - \big(E(X)\big)^2 = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2.$$

Result:

If $X \sim$ Gamma(α, β), then

$$\mu_X = E(X) = \alpha\beta \quad \textit{and}$$

$$\sigma_X^2 = Var(X) = \alpha\beta^2.$$

Moment-generating function of $X \sim$ Gamma(α, β).

$$\begin{aligned}
M(\theta) &= E(e^{\theta X}) \\
&= \int_0^\infty e^{\theta x} \cdot \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}\, dx \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x/\beta + \theta x}\, dx \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty \underbrace{x^{\alpha-1} e^{-x\left(\frac{1}{\beta}-\theta\right)}}_{\substack{\text{shape}=\alpha \\ \text{scale}=\left(\frac{1}{\beta}-\theta\right)^{-1}}} dx \qquad (\text{need } \theta < \frac{1}{\beta}) \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \left(\left(\frac{1}{\beta} - \theta \right)^{-1} \right)^\alpha \Gamma(\alpha) \\
&= (1 - \beta\theta)^{-\alpha}.
\end{aligned}$$

Result:
If $X \sim \text{Gamma}(\alpha, \beta)$, then

$$M_X(\theta) = (1 - \beta\theta)^{-\alpha}.$$

Example.
Use this MGF to compute the first two moments of the Gamma(α, β).

SOLUTION:
From the properties of the MGF on page 131 Fact 3, we can take derivatives of the MGF with respect to its argument θ and then evaluate at $\theta = 0$ to recover the moments of the distribution.

$M(\theta) = (1 - \beta\theta)^{-\alpha}$

$M'(\theta) = -\alpha \cdot (1 - \beta\theta)^{-\alpha-1} \cdot (-\beta) = \alpha\beta(1 - \beta\theta)^{-(\alpha+1)}$
$\implies M'(0) = \alpha\beta = E(X).$

$M''(\theta) = -(\alpha + 1)\alpha\beta(1 - \beta\theta)^{-(\alpha+2)}(-\beta) = \alpha(\alpha + 1)\beta^2(1 - \beta\theta)^{-(\alpha+2)}$
$\implies M''(0) = \alpha(\alpha + 1)\beta^2 = E(X^2).$

The Normal distribution (a.k.a. the Gaussian distribution).

$X \sim \mathrm{N}(\mu, \sigma^2)$ means the continuous rv X has the pdf

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \quad \text{for } -\infty < x < \infty.$$

The parameters in this pdf, μ and σ^2, will end up being the mean and the variance, respectively – so it is only natural to use these symbols for the parameters (see the remark on page 165). This pdf is one of a few pdfs in this course whose support is the *entire* real line. Another we've talked about was the *Cauchy distribution* (page 145).

Like the Gamma(α, β) we'll be interested in the global behavior of these pdfs; However, unlike the Gamma(α, β), it will be quite important to be able to find areas under these pdfs – and this will require the use of a statistical table or a computing device programmed to find these areas – and, this will come soon.

Here are some plots[4] of the $\mathrm{N}(\mu, \sigma^2)$ pdf for various choices of μ and σ^2:

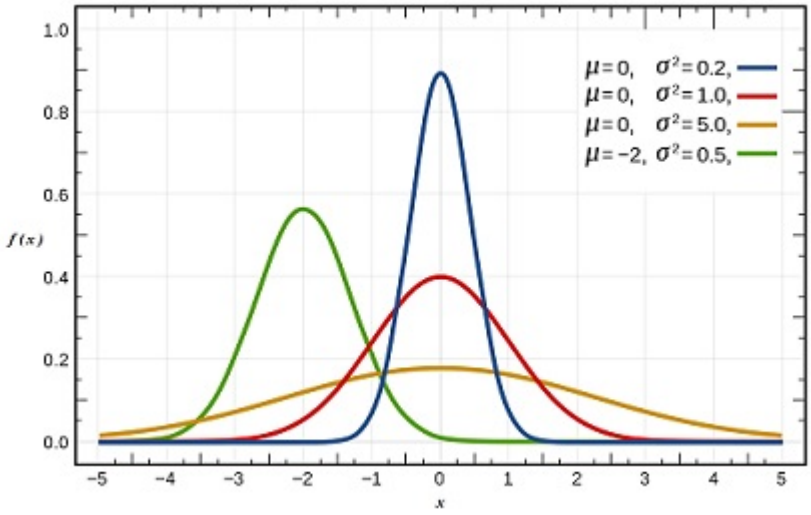


Figure. Some plots of Normal pdfs; the red one is the *standard normal pdf*.

Remark.
The Normal distribution is important for several reasons, but largely because it happens to be (roughly speaking) the limiting distribution of large sums of independent random variables. We'll see, for instance, that

- for *fixed* p and *large* n, binom$(n, p) \approx \mathrm{N}(np, np(1-p))$,
- for *large* λ, Poisson$(\lambda) \approx \mathrm{N}(\lambda, \lambda)$,
- for *fixed* β and *large* α, Gamma$(\alpha, \beta) \approx \mathrm{N}(\alpha\beta, \alpha\beta^2)^{\dagger}$,

and *much much more!*

† Indeed, look at the Gamma densities with $\alpha = 9$ and $\alpha = 7.5$ on page 155 and you can see these are already starting to look like Normal pdfs!

[4]Figure modified from https://en.wikipedia.org/wiki/Normal_distribution.

Remark.
We should probably show that the Normal pdf given on page 159 is *actually* a pdf; i.e., it's *nonnegative* on the entire real line *and* the total integral is 1. The clever idea in the proof is due to C.F. Gauss and is one of the reasons this distribution is often called the ***Gaussian distribution***.

Proof that the $\mathbf{N}(\mu, \sigma^2)$ pdf *is* a pdf.
Since $e^u > 0$ for any real u, $f(x) > 0$ for $x \in \mathbb{R}$. To show $\mathcal{I} := \int_{-\infty}^{\infty} f(x)\,dx = 1$, Gauss showed that $\mathcal{I}^2 = 1$, from which it follows $\mathcal{I} = \pm 1$ and we can rule out -1 since $f(x) > 0$ everywhere.

$$
\begin{aligned}
\mathcal{I}^2 &= \left(\int_{-\infty}^{\infty} f(x)\,dx\right)^2 \\
&= \int_{-\infty}^{\infty} f(x)\,dx \int_{-\infty}^{\infty} f(y)\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)f(y)\,dxdy \\
&= \frac{1}{2\pi\sigma^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\cdot e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}\,dxdy \quad \left(u = \frac{x-\mu}{\sigma},\ v = \frac{y-\mu}{\sigma}\right) \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\cdot e^{-\frac{1}{2}v^2}\,dudv \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(u^2+v^2)}\,dudv \quad \left(\text{polar coordinates: } r^2 = u^2+v^2,\ dudv = rdrd\theta\right) \\
&= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} re^{-\frac{r^2}{2}}\,drd\theta \\
&= \frac{1}{2\pi}\int_{0}^{2\pi}\left(-e^{-\frac{r^2}{2}}\Big|_{r=0}^{\infty}\right)\,d\theta \\
&= \frac{1}{2\pi}\int_{0}^{2\pi} 1\,d\theta \\
&= 1.
\end{aligned}
$$

As we did with the Gamma(α, β) distribution, we will first focus on deriving *global* properties of the Normal distribution.

Properties of the $\mathrm{N}(\mu, \sigma^2)$ distribution.

Theorem. (An affine transformation of a Normal is, again, a Normal)
Let $X \sim \mathrm{N}(\mu, \sigma^2)$, and let a, b be any constants with $a \neq 0$. Then

$$aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2).$$

Remark.
The theorem is also true when $a = 0$ if we interpret the $\mathrm{N}(a\mu + b, a^2\sigma^2) = \mathrm{N}(b, 0)$ distribution as a point mass of 1 centered at b. There are several techniques that probabilists use to find the distribution of functions of continuous random variable(s), but for the above theorem I will foreshadow a bit and apply the ***CDF method***.

Outline of the CDF method.[5]
Suppose $Y = g(X)$, where X is a continuous rv having *known* CDF $F_X(x)$. The main idea in this method is to find the CDF of Y in terms of the CDF of X, then once we've found the CDF of Y, we know its derivative is the pdf of Y.

Proof.
Let $Y = aX + b$ and first assume $a > 0$. The case $a < 0$ will be an exercise.
Since X can take on any value in $\mathbb{R}$, it follows Y can take on any value in $\mathbb{R}$, too, because $a \neq 0$. So, let $y \in \mathbb{R}$ be arbitrary.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(aX + b \leq y) \\ &= P(aX \leq y - b) \\ &= P\left(X \leq \frac{y-b}{a}\right) \qquad \text{(notice I used the fact that } a > 0 \text{ here)} \\ &= F_X\left(\frac{y-b}{a}\right), \end{aligned}$$

and we've found the CDF of Y in terms of the CDF of X. Now, if we take the derivalive on both sides with respect to y, then on the left we get the pdf of Y, and on the right we get (after applying the chain rule):

$$\begin{aligned} f_Y(y) &= f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{\left(\frac{y-b}{a} - \mu\right)^2}{2\sigma^2}} \cdot \frac{1}{a} \\ &= \frac{1}{\sqrt{2\pi a^2\sigma^2}}\, e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}}, \end{aligned}$$

which is the pdf of a $\mathrm{N}(a\mu + b, a^2\sigma^2)$ which was to be shown.

[5] we will learn the CDF method in more detail later. See page 172.

Remark.
In the proof we found that $F_Y(y) = P\left(X \le \frac{y-b}{a}\right)$, which can be written equivalently as

$$F_Y(y) = \int_{-\infty}^{\frac{y-b}{a}} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\, dx.$$

Integrals whose bounds (and possibly integrand) involve the variable that we wish to take derivative against come up often in probability, and the calculus rule we appeal to is called ***the Liebniz rule***.

Calculus fact: Liebniz rule.

$$\frac{d}{dy}\int_{a(y)}^{b(y)} g(x,y)\, dx = g(b(y),y)b'(y) - g(a(y),y)a'(y) + \int_{a(y)}^{b(y)} \frac{\partial g(x,y)}{\partial y}\, dx.$$

Some special cases of the Liebniz rule:

$g(x,y) = g(x)$ does not depend on y: $\frac{d}{dy}\int_{a(y)}^{b(y)} g(x)\, dx = g(b(y))b'(y) - g(a(y))a'(y)$.

$a(y) = a$ is constant (incl. $-\infty$), no y-dependence in g: $\frac{d}{dy}\int_{a}^{b(y)} g(x)\, dx = g(b(y))b'(y)$.

Both bounds are constant: $\frac{d}{dy}\int_a^b g(x,y)\, dx = \int_a^b \frac{\partial g(x,y)}{\partial y}\, dx$.

We can apply the Liebniz rule on

$$F_Y(y) = \int_{-\infty}^{\frac{y-b}{a}} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\, dx$$

to recover the pdf of Y. Let's do this now:

$$\begin{aligned}
\frac{d}{dy}F_Y(y) &= \frac{d}{dy}\int_{-\infty}^{\frac{y-b}{a}} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\, dx \\
&= \frac{e^{-\frac{(\frac{y-b}{a}-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\,\sigma^2}} \cdot \underbrace{\frac{d}{dy}\left(\frac{y-b}{a}\right)}_{=\frac{1}{a}} \\
&= \frac{1}{\sqrt{2\pi a^2\sigma^2}}\, e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}},
\end{aligned}$$

which confirms (in an equivalent way) what we did on the previous page.

Exercise for the student.
If $X \sim \mathrm{N}(\mu, \sigma^2)$ and a, b constants with $a < 0$, show that $Y = zX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$.
Hint: Because $a < 0$, we find that, for real y,

$$F_Y(y) = P(aX \le y - b) = P(X \ge \frac{y-b}{a}) = \int_{\frac{y-b}{a}}^{\infty} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\, dx = 1 - \int_{-\infty}^{\frac{y-b}{a}} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\, dx.$$

Now use the Liebniz rule.

After you prove the exercise, we will have shown that, if $X \sim \mathrm{N}(\mu, \sigma^2)$, for any constants a, b with $a \neq 0$,

$$aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2).$$

We have the immediate corollaries:

Corollary.
(a) If $X \sim \mathrm{N}(\mu, \sigma^2)$, then the ***z-score*** $Z := \frac{X-\mu}{\sigma} \sim \mathrm{N}(0, 1)$.

(b) Conversely, if $Z \sim \mathrm{N}(0, 1)$, then, for any $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, $\mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$.

Proof.
(a) Take $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$. Then $a\mu + b = \frac{1}{\sigma}\mu + \left(-\frac{\mu}{\sigma}\right) = 0$, and $a^2\sigma^2 = \left(\frac{1}{\sigma}\right)^2 \sigma^2 = 1$.

(b) $Z \sim \mathrm{N}(0, 1)$, so $aZ + \mu \sim \mathrm{N}(\sigma \cdot 0 + \mu, \sigma^2(1)^2) = \mathrm{N}(\mu, \sigma^2)$.

Remark.
We call the $\mathrm{N}(0, 1)$ the ***standard Normal distribution***. Its pdf is

$$\varphi(x) = \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} \quad \text{for } -\infty < x < \infty,$$

and its CDF is

$$\Phi(x) := \int_{-\infty}^{x} \varphi(u)\, du.$$

The Greek letters φ and Φ are *reserved* letters for the standard Normal pdf and CDF, respectively. Also, for standard Normal rvs we will typically use the capital letter Z or $Z_1, Z_2, \ldots$, etc.

The corollary clearly shows that we can transform any Normal distribution into a ***standard Normal distribution***, i.e., a $\mathrm{N}(0, 1)$ distribution, and vice-versa. Thus, any probability involving a $\mathrm{N}(\mu, \sigma^2)$ rv can be transformed into a probability involving a standard normal rv. Therefore, if we tabulate the CDF of a standard Normal rv we can use this *one table* to tabulate the CDF of every Normal rv.

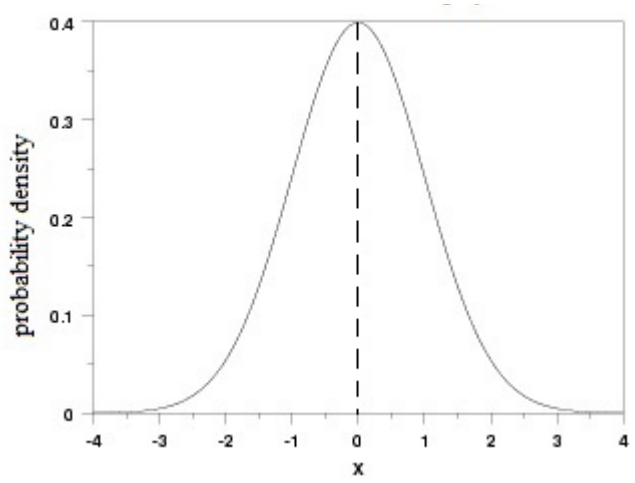


Figure. Plot of the standard normal pdf $\varphi(x)$. See also the red curve on page 159.

Symmetry of the Normal pdfs.

Notice $\varphi(-x) = \dfrac{e^{-\frac{1}{2}(-x)^2}}{\sqrt{2\pi}} = \dfrac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} = \varphi(x)$, which implies the standard Normal distribution is a ***symmetric distribution***, in fact, it is symmetric about $x = 0$.
Do you see why this implies $\Phi(0) = \frac{1}{2}$ *?*

Exercise for the student.

Let $Z \sim \mathrm{N}(0, 1)$.
(a) Show that $E(Z)$ exists. Then show $E(Z) = 0$.
(b) Show $E(Z^2) = 1$.
(c) If $X \sim \mathrm{N}(\mu, \sigma^2)$, then explain why the pdf of X is also symmetric (about $x = \mu$, in fact).

SOLUTION to part (b):
Since $Z^2 \geq 0$ there is no issue with $E(Z^2)$ not existing; it may be infinite but we can just compute it to see.

$$\begin{aligned} E(Z^2) &= \int_{-\infty}^{\infty} z^2 \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}\, dz \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} z \left(z\, e^{-\frac{1}{2}z^2} \right) dz \qquad \left(\begin{array}{ll} u = z, & dv = z e^{-\frac{1}{2}z^2}\, dz \\ du = dz & v = -e^{-\frac{1}{2}z^2} \end{array} \right) \\ &= \frac{2}{\sqrt{2\pi}} \left(\underbrace{-z e^{-\frac{1}{2}z^2} \Big|_{z=0}^{\infty}}_{=0} \right) + \underbrace{\frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}z^2}\, dz}_{=1} \\ &= 1. \end{aligned}$$

Since $E(Z^2) = 1$ and (you showed) $E(Z) = 0$, it follows $Var(Z) = E(Z^2) - \big(E(Z)\big)^2 = 1$.

Remark. (In a $\mathrm{N}(\mu, \sigma^2)$, μ is the mean and σ^2 is the variance)
By the corollary on page 163 ($Z \sim \mathrm{N}(0,1) \implies X = \sigma Z + \mu \sim \mathrm{N}(\mu, \sigma^2)$) and, by the property of expected value on page 121,

$$E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \sigma \cdot 0 + \mu = \mu,$$

and, by the property of variance on page 129,

$$Var(X) = Var(\sigma Z + \mu) = \sigma^2 Var(Z) = \sigma^2.$$

Therefore, the two parameters defining the $\mathrm{N}(\mu, \sigma^2)$ distribution *are* ***the mean*** and ***the variance***.

Moment-generating function of a standard Normal.
Let $Z \sim \mathrm{N}(0,1)$. Then the MGF of Z is defined for *all* real θ and

$$M_Z(\theta) = E(e^{\theta Z}) = e^{\frac{\theta^2}{2}}.$$

Proof.

$$\begin{aligned}
M(\theta) &= E(e^{\theta Z}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\theta z} \cdot e^{-\frac{1}{2}z^2}\, dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2 - 2\theta z)}\, dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2 - 2\theta z + \theta^2) + \frac{1}{2}\theta^2}\, dz \quad \text{(completed the square)} \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-\theta)^2} e^{\frac{1}{2}\theta^2}\, dz \\
&= \frac{e^{\frac{\theta^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-\theta)^2}\, dz \qquad (u = z - \theta,\ du = dz) \\
&= e^{\frac{\theta^2}{2}} \underbrace{\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}u^2}}{\sqrt{2\pi}}\, du}_{=1} \quad = \quad e^{\frac{\theta^2}{2}}.
\end{aligned}$$

Moments of the standard Normal.
Using the same idea we used to get a general formula for the moments of an $\exp(\lambda)$ (see page 149), we can find a general formula for the moments of a standard normal.
The MacLaurin expansion for $M_Z(\theta)$ is

$$\begin{aligned}
e^{\frac{\theta^2}{2}} &= 1+\left(\frac{\theta^2}{2}\right)+\frac{1}{2!}\left(\frac{\theta^2}{2}\right)^2+\frac{1}{3!}\left(\frac{\theta^2}{2}\right)^2+\frac{1}{4!}\left(\frac{\theta^2}{2}\right)^2+\cdots \\
&= 1+\frac{\theta^2}{2}+\frac{\theta^4}{2!2^2}+\frac{\theta^6}{3!2^3}+\frac{\theta^8}{4!2^4}+\cdots \\
&= 1+\frac{\theta^2}{2!}+\frac{4!}{2!2^2}\cdot\frac{\theta^4}{4!}+\frac{6!}{3!2^3}\cdot\frac{\theta^6}{6!}+\frac{8!}{4!2^4}\cdot\frac{\theta^8}{8!}+\cdots \\
&= 1+\underbrace{0}_{=E(Z)}\cdot\theta+\underbrace{1}_{=E(Z^2)}\cdot\frac{\theta^2}{2!}+\underbrace{0}_{=E(Z^3)}\cdot\frac{\theta^3}{3!}+\underbrace{\frac{4!}{2!2^2}}_{=E(Z^4)}\cdot\frac{\theta^4}{4!}+\underbrace{0}_{=E(Z^5)}\cdot\frac{\theta^5}{5!}+\underbrace{\frac{6!}{3!2^3}}_{=E(Z^6)}\cdot\frac{\theta^6}{6!}+\cdots
\end{aligned}$$

from which it follows the odd moments of a standard Normal vanish and the even moments are $(2k)!/[k!2^k]$: For $k=1,2,3,\ldots,$

$$E(Z^{2k-1})=0$$

and

$$E(Z^{2k})=\frac{(2k)!}{k!2^k}.$$

MGF of a $\mathrm{N}(\mu,\sigma^2)$.
From the corollary on page 163, once we have the MGF for a standard Normal we also have it for a $\mathrm{N}(\mu,\sigma^2)$, since $Z\sim\mathrm{N}(0,1)\implies\sigma Z+\mu\sim\mathrm{N}(\mu,\sigma^2)$: If $X\sim\mathrm{N}(\mu,\sigma^2)$, then

$$M_X(\theta)=E(e^{\theta X})=E(e^{\theta(\sigma Z+\mu)})=E(e^{\theta\sigma Z}e^{\theta\mu})=e^{\theta\mu}\underbrace{E(e^{(\theta\sigma)Z})}_{=M_Z(\theta\sigma)}=e^{\theta\mu}e^{\frac{\theta^2\sigma^2}{2}}$$

So, if $X\sim\mathrm{N}(\mu,\sigma^2)$, then

$$M_X(\theta)=e^{\theta\mu+\frac{\theta^2\sigma^2}{2}}.$$

Computing probabilities involving a standard Normal rv.
When Z is standard Normal, i.e., $Z \sim \mathrm{N}(0,1)$, we have the CDF of Z:

$$\Phi(x) = \int_{-\infty}^{x} \frac{e^{-\frac{1}{2}u^2}}{\sqrt{2\pi}}\, du,$$

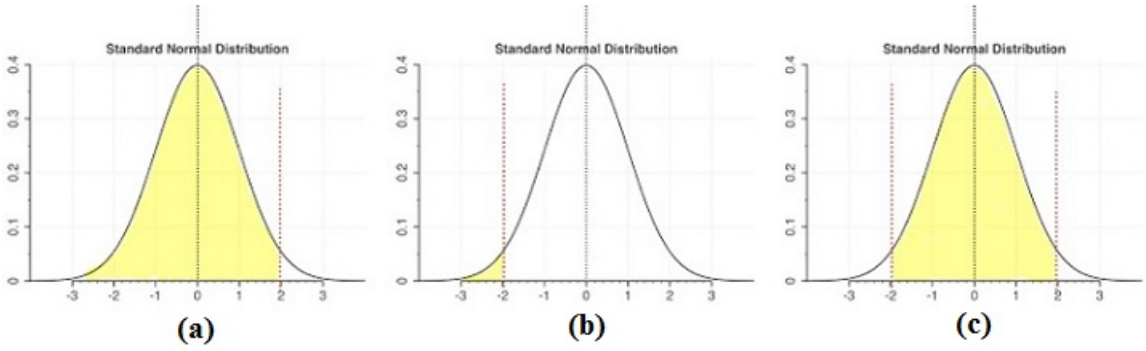
and on page 160 we showed $\Phi(\infty) = 1$, and on page 164 we showed $\Phi(0) = \frac{1}{2}$. But, unfortunately, these are the only values where Φ can be computed exactly, and, if we *needed* the value of $\Phi(x)$ for any other x we can only approximate it.

Many calculators and computers now have the functionality to give the value of $\Phi(x)$ for any desired x to many decimal-place accuracy. By the way, I'll point out that once we can approximate $\Phi(x)$, we can also approximate the CDF of *any* Normal rv $X \sim \mathrm{N}(\mu, \sigma^2)$ by

$$F_X(x) = P(X \le x) = P(\sigma Z + \mu \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

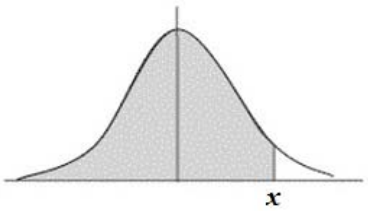
Since it will be necessary to evaluate probabilities involving Normal rvs, I'll present a statistical table with approximate values for $\Phi(x)$, where x will range from -3.49 to 3.49 in increments of .01.

Examples.
Verify the following using the standard Normal table on page 168:
(a) $P(Z \le 1.96) = .9750$
(b) $P(Z \le -1.96) = .0250$
(c) $P(-1.96 \le Z \le 1.96) = .9500$



(a) (b) (c)

By the symmetry of the standard Normal pdf, it should be clear that $P(Z \ge 1.96) = P(Z \le -1.96) = .0250$, this can also be seen by the complementary rule: $P(Z \ge 1.96) = 1 - P(Z < 1.96) = 1 - .9750 = .0250$.



$$\Phi(x) = P(Z < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du$$

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Computing probabilities with arbitrary Normals.
Now that we know how to use a standard Normal table, how can it be used to compute probabilities involving an arbitrary $X \sim \mathrm{N}(\mu, \sigma^2)$ rv?

Recall the CDF of X is $F_X(x) = \Phi\big(\frac{x-\mu}{\sigma}\big)$, i.e., $\frac{X-\mu}{\sigma} = Z$ is a standard Normal. Let's do an example:

Example.
Assume the adult male height X (in inches) of a certain population is normally distributed with a mean of 70 inches and a standard deviation of $\sigma = 4$ inches (i.e, variance $\sigma^2 = 16$ inches^2) – a long-winded way of saying $X \sim \mathrm{N}(70, 16)$.
(a) What proportion of the population is taller than 76 inches?
(b) What proportion of the population is between 66 and 74 inches?

SOLUTION:
(a) $P(X > 76) = P\big(\frac{X-\mu}{\sigma} > \frac{76-70}{4}\big) = P(Z > 1.50) = 1 - P(Z \leq 1.50) = 1 - .9332 = .0668$; approximately 6.68% of the population is taller than 76 inches.
(b) $P(66 \leq X \leq 74) = P\big(\frac{66-70}{4} \leq \frac{X-\mu}{\sigma} \leq \frac{74-70}{4}\big) = P(-1.00 \leq Z \leq 1.00) = P(Z \leq 1.00) - P(Z \leq -1.00) = .8413 - .1587 = .6826$; approximately 68.26% of the population is between 66 and 74 inches.

Extreme behavior of random variables. (***This is a digression.***)

Sometimes it is important to understand the extreme behavior of random variables, i.e., the probability the rv takes extremely large (or small) values. One particular instance where this is the case is in ***Reliability theory***.

Let X be a continuous rv with pdf $f(x)$ and CDF $F(x)$.

We define the ***survival function***:

$$\overline{F}(x) := P(X > x),$$

as you can see $\overline{F}(x) := 1 - F(x)$. If X represents the lifetime of a component (time to death), then $\overline{F}(x)$ is the probability the component is alive at x – it has survived to time x.

We define the ***hazard rate function***:

$$\begin{aligned} h(x) &:= \lim_{\delta \downarrow 0} \frac{P(x < X \leq x + \delta | X > x)}{\delta} \\ &= \lim_{\delta \downarrow 0} \frac{P((x < X \leq x + \delta) \cap (X > x))}{\delta P(X > x)} \\ &= \lim_{\delta \downarrow 0} \frac{P(x < X \leq x + \delta)}{\delta P(X > x)} \\ &= \lim_{\delta \downarrow 0} \frac{F(x + \delta) - F(x)}{\delta \overline{F}(x)} = \frac{f(x)}{\overline{F}(x)}. \end{aligned}$$

The ***Mills ratio*** is the reciprocal of the hazard rate function $h(x)$:

$$m(x) = \frac{\overline{F}(x)}{f(x)} = \frac{P(X > x)}{f(x)}.$$

Probabilists and Statisticians are often interested in large x behavior of these quantities.

Notation.
We write $a(x) \sim b(x)$ as $x \to \infty$ to mean

$$\lim_{x \to \infty} \frac{a(x)}{b(x)} = 1.$$

Example.
Let $Z \sim \mathrm{N}(0,1)$. Show that $m(x) \sim \frac{1}{x}$ as $x \to \infty$.
A consequence of this is that $m(x) := \frac{P(Z>x)}{\varphi(x)} \sim \frac{1}{x} \implies P(Z > x) \approx \frac{\varphi(x)}{x}$ as $x \to \infty$.

SOLUTION:

$$\begin{aligned} P(Z > x) &= \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-z^2/2}\, dz \\ &= \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{1}{z} \left(z e^{-z^2/2} \right) dz \\ &= \frac{1}{\sqrt{2\pi}} \left(-\frac{1}{z} e^{-z^2/2} \Big|_{z=x}^\infty - \int_x^\infty \frac{1}{z^2} e^{-z^2/2}\, dz \right) \\ &= \frac{1}{\sqrt{2\pi}} \left(\frac{1}{x} e^{-x^2/2} - \int_x^\infty \frac{1}{z^2} e^{-z^2/2}\, dz \right) \\ &= \frac{1}{x}\varphi(x) - \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{1}{z^2} e^{-z^2/2}\, dz. \end{aligned}$$

Therefore,

$$m(x) = \frac{P(Z > x)}{\varphi(x)} = \frac{1}{x} - \frac{\frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{1}{z^2} e^{-z^2/2}\, dz}{\varphi(x)},$$

and, using L'Hôpital's rule, the last term tends of 0 as $x \to \infty$ (please check this for yourself), and then this proves that

$$\lim_{x \to \infty} \frac{P(Z > x)}{\frac{1}{x}\varphi(x)} = 1,$$

which basically says that, for *large* x, $P(Z > x)$ can be well-approximated by $\frac{\varphi(x)}{x}$. In fact, the approximation is already rather good for "small" x, let's see....

From the table...	$m(x)\varphi(x) = \frac{\varphi(x)}{x}$	Abs.rel.error$= \left\lvert \frac{P(Z>x) - \frac{\varphi(x)}{x}}{\frac{\varphi(x)}{x}} \right\rvert$
$P(Z > 1) = .1587$	.241971	.34
$P(Z > 2) = .0228$	.026995	.16
$P(Z > 3) = .0013$	.001477	.086
$P(Z > 4)^* = .00003167$	.00003345	.053

Figure. * used software to find this value.

One can repeat what we did with the standard Normal above *ad infinitum* to show that the Mills ratio has the expansion $m(x) = \frac{1}{x} - \frac{1}{x^3} + \frac{1}{x^5} - + \cdots$.

The CDF method – univariate case.

The CDF method is one of several approaches[6] we will learn that solves the following ***general problem:***

Suppose X is a *continuous* rv with pdf $f_X(x)$. Find the pdf of $Y = g(X)$.

In fact, the CDF method can also handle the ***multivariate case*** where we have *more than one* jointly continuous rvs $X_1, X_2, \ldots, X_n$ and we wish to find the pdf of some function of them: $Y = g(X_1, X_2, \ldots, X_n)$. We will demonstrate this *after* we've introduced the concept of jointly distributed rvs. But, in this section, we will concentrate on the ***univariate case***, i.e., the case involving only one random variable.

We've already used this method back on page 161 to show when X has the pdf

$$f_X(x) = \frac{e^{-\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \quad \text{for } -\infty < x < \infty,$$

then $Y = g(X) := aX + b$ has the pdf

$$f_Y(y) = \frac{e^{-\frac{(x-(a\mu+b))^2}{2a^2\sigma^2}}}{\sqrt{2\pi a^2\sigma^2}} \quad \text{for } -\infty < y < \infty.$$

Steps to apply in the CDF method.

step 1. Identify the image of $Y = g(X)$, i.e., the support of $f_Y(y)$.

Since $f_Y(y) \equiv 0$ for y not in the support, we can concentrate solely on $y \in \text{supp}(f_Y)$.

step 2. For $y \in \text{supp}(f_Y)$, compute the CDF of Y: $F_Y(y) := P(Y \leq y) = P(g(X) \leq y)$.

Here we need to manipulate the CDF $P(g(X) \leq y)$ into one that involves only the CDF of X.

step 3. Once we have the CDF of Y in terms of the CDF of X, take the derivative:

We may need to use a chain rule or a Liebniz rule to take this derivative.
$f_Y(y) = \frac{d}{dy}F_Y(y)$ for $y \in \text{supp}(f_Y(y))$,
$f_Y(y) = 0$ for $y \notin \text{supp}(f_Y(y))$.

[6]other approaches we'll learn in this course: method of Jacobians, convolution, and MGF method.

Example.
Suppose $Z \sim \mathrm{N}(0,1)$. Find the pdf of $Y = Z^2$.

SOLUTION:
Here, $Y = g(Z)$, where $g(z) = z^2$. Since a standard Normal rv takes values from $-\infty < z < \infty$, it follows that $y = g(z) = z^2$ takes values from $0 \le y < \infty$. Thus, $\mathrm{supp}(f_Y) = [0, \infty)$. Right away, this tells us that $f_Y(y) = 0$ for $y \le 0$, and we can concentrate on $y > 0$ (don't forget: Y is also a continuous rv, so $P(Y = 0) = 0$).

Now, let $y > 0$.

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(Z^2 \le y) \\ &= P(|Z| \le \sqrt{y}) \qquad \text{(remember! } \sqrt{Z^2} = |Z|\text{.)} \\ &= P(-\sqrt{y} \le Z \le \sqrt{y}) \\ &= \int_{-y^{\frac{1}{2}}}^{y^{\frac{1}{2}}} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz. \end{aligned}$$

At this point you can either recognize that this integral is really just $\Phi(y^{\frac{1}{2}}) - \Phi(-y^{\frac{1}{2}})$ (and then you can take the derivative in y not forgetting to use the chain rule), or, you can (equivalently) just use the Liebniz rule to take the derivative in y which I do now:

$$\begin{aligned} f_Y(y) = \frac{d}{dy}\int_{-y^{\frac{1}{2}}}^{y^{\frac{1}{2}}} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz &= \frac{e^{-\left(y^{\frac{1}{2}}\right)^2/2}}{\sqrt{2\pi}} \cdot \frac{d}{dy}\left(y^{\frac{1}{2}}\right) - \frac{e^{-\left(-y^{\frac{1}{2}}\right)^2/2}}{\sqrt{2\pi}} \cdot \frac{d}{dy}\left(-y^{\frac{1}{2}}\right) \\ &= \frac{e^{-y/2}}{\sqrt{2\pi}} \cdot \underbrace{\frac{d}{dy}\left(y^{\frac{1}{2}}\right)}_{=\frac{y^{-\frac{1}{2}}}{2}} + \frac{e^{-y/2}}{\sqrt{2\pi}} \cdot \underbrace{\frac{d}{dy}\left(y^{\frac{1}{2}}\right)}_{=\frac{y^{-\frac{1}{2}}}{2}} \\ &= \frac{y^{-\frac{1}{2}}e^{-y/2}}{2^{\frac{1}{2}}\sqrt{\pi}} = \frac{y^{\frac{1}{2}-1}e^{-y/2}}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})}, \end{aligned}$$

which *hopefully* you recognize as the pdf of a Gamma$(\frac{1}{2}, 2)$.

Thus, we've shown

$$f_Y(y) = \begin{cases} \dfrac{y^{\frac{1}{2}-1}e^{-y/2}}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} & \text{for } y > 0 \\ \\ 0 & \text{for } y \le 0 \end{cases}.$$

Statisticians call this distribution the ***Chi-square with 1 degree of freedom*** – it is the distribution of the square of a standard Normal.

Strictly monotone transformations of a continuous rv.

The CDF method is used to prove the following theorem:

Theorem.
Suppose X is a continuous rv with pdf $f_X(x)$ and $Y = g(X)$ with $g = g(x)$ being strictly monotone. Then

$$f_Y(y) = f_X\big(g^{-1}(y)\big) \cdot \left| \frac{dg^{-1}(y)}{dy} \right|.$$

Math facts.
A function $g = g(x)$ is strictly increasing (resp., strictly decreasing) means if $a < b$ then $g(a) < g(b)$ (resp., $g(a) > g(b)$); and, we call such functions ***strictly monotone***.
FACT: If $g = g(x)$ is strictly increasing (resp., strictly decreasing), then $g^{-1} = g^{-1}(y)$ exists and is strictly increasing (resp., strictly decreasing).

Proof.
Let's first prove this in the case that $g = g(x)$ is strictly increasing.
Let $I = \text{supp}(f_X)$, then $g(I) = \{y : y = g(x) \text{ for some } x \in \text{supp}(f_X)\} = \text{supp}(f_Y)$. Take any $y \in \text{supp}(f_Y)$.

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(g(X) \le y) \\ &= P(X \le g^{-1}(y)) \quad (g \text{ increasing } \implies g(X) \le y \implies g^{-1}(g(X)) \le g^{-1}(y)) \\ &= F_X(g^{-1}(y)). \end{aligned}$$

Thus, $F_Y(y) = F_X(g^{-1}(y))$. Now, since the derivative of the CDF is the pdf, we have

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= \frac{dF_X(g^{-1}(y))}{dy} = f_X(g^{-1}(y)) \cdot \frac{g^{-1}(y)}{dy}, \quad (g^{-1} \text{ increasing } \implies \frac{dg^{-1}(y)}{dy} > 0) \end{aligned}$$

where we used the chain rule in the last step. Alternatively, we could've written

$$F_Y(y) = \int_{-\infty}^{g^{-1}(y)} f_X(x)\, dx,$$

and appealed to the Liebniz rule.

Suppose, instead, $g = g(x)$ is strictly decreasing.
Then

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(g(X) \le y) \\ &= P(X \ge g^{-1}(y)) \quad (g \text{ decreasing } \implies g(X) \le y \implies g^{-1}(g(X)) \ge g^{-1}(y)) \\ &= \int_{g^{-1}(y)}^{\infty} f_X(x)\,dx, \end{aligned}$$

and appealing to the Liebniz rule,

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}\int_{g^{-1}(y)}^{\infty} f_X(x)\,dx \\ &= -f_X(g^{-1}(y)) \cdot \frac{dg^{-1}(y)}{dy} \\ &= f_X(g^{-1}(y)) \cdot \frac{d\left(-g^{-1}(y)\right)}{dy}. \quad (g^{-1} \text{ decreasing } \implies \frac{d\left(-g^{-1}(y)\right)}{dy} > 0) \end{aligned}$$

Since we get a pdf for Y regardless of whether $g = g(x)$ strictly increasing or strictly decreasing, we can write the resulting pdf for Y as:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{dg^{-1}(y)}{dy}\right|.$$

□

Example.
Let's show, again, that if $X \sim \mathrm{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$.

SOLUTION:
$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$z = g(x) = \frac{x-\mu}{\sigma} \implies g^{-1}(z) = \sigma z + \mu \implies \frac{d}{dz}\left(g^{-1}(z)\right) = \sigma.$

$$\begin{aligned} f_Z(z) &= f_X(g^{-1}(z)) \cdot \left|\frac{d}{dz}\left(g^{-1}(z)\right)\right| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\sigma z+\mu-\mu)^2}{2\sigma^2}} \cdot |\sigma| \quad (|\sigma| = \sigma \text{ since } \sigma > 0) \\ &= \frac{\sigma}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{z^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}} = \varphi(z), \end{aligned}$$

and we see the z-score of a Normal is a standard Normal.

Remark.
The z-score of a Normal rv will still have a Normal distribution. This isn't necessarily true of other distributions. For example, the z-score of a Gamma(α, β) will *not* have a Gamma distribution as the next example demonstrates.

Example.
Let $X \sim$ Gamma(α, β). Then the mean of X is $\mu = \alpha\beta$ and the variance of X is $\sigma^2 = \alpha\beta^2$, i.e., $\sigma = \beta\sqrt{\alpha}$. Find the pdf of $Z = \frac{X-\alpha\beta}{\beta\sqrt{\alpha}}$ and show that it is *not* Gamma distributed.

SOLUTION:
For $x > 0$, $f_X(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$.
$z = \frac{x-\alpha\beta}{\beta\sqrt{\alpha}} \implies z > -\sqrt{\alpha}$ when $x > 0$. So, it appears that supp(f_Z) is *not* $[0, \infty)$! Not a good sign if we are hoping for a Gamma!!

$g^{-1}(z) = \beta\sqrt{\alpha}z + \alpha\beta \implies \frac{d}{dz}(g^{-1}(z)) = \beta\sqrt{\alpha} > 0$.
So, for $z > -\sqrt{\alpha}$,

$$\begin{aligned} f_Z(z) &= f_X(\beta\sqrt{\alpha}z + \alpha\beta) \cdot \beta\sqrt{\alpha} \\ &= \frac{(\beta\sqrt{\alpha}z + \alpha\beta)^{\alpha-1}e^{-(\beta\sqrt{\alpha}z+\alpha\beta)/\beta}}{\beta^\alpha\Gamma(\alpha)} \\ &= \frac{\beta^{\alpha-1}(\sqrt{\alpha})^{\alpha-1}(z+\sqrt{\alpha})^{\alpha-1}e^{-(\sqrt{\alpha}z+\alpha)}}{\beta^\alpha\Gamma(\alpha)} \\ &= \frac{\alpha^{\frac{\alpha-1}{2}}(z+\sqrt{\alpha})^{\alpha-1}e^{-\sqrt{a}(z+\sqrt{\alpha})}}{\Gamma(\alpha)} \end{aligned}$$

and this pdf is positive, for instance, when $z = -\frac{1}{2}\sqrt{\alpha}$ and, thus, it's support is not the nonnegative reals *and*, therefore, cannot be Gamma distributed.

Exercise for the student.
Suppose $X \sim \exp(1)$, i.e., X is the continuous rv with pdf

$$f(x) = e^{-x} \quad \text{for } x > 0,$$

sometimes called the ***unit exponential distribution***. Fix $\nu, \alpha, \beta \in \mathbb{R}$ with $\alpha > 0$ and $\beta > 0$. Find the pdf of

$$Y = \nu + \alpha X^{\frac{1}{\beta}}.$$

The distribution of Y is called the ***Weibull distribution***.

ANSWER:

$$f_Y(y) = \begin{cases} \frac{\beta(y-\nu)^{\beta-1}}{\alpha^\beta} e^{-\left(\frac{y-\nu}{\alpha}\right)^\beta} & \text{for } y > \nu \\ 0 & \text{for } y \leq \nu \end{cases}.$$

Example. (the log-normal distribution) Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$. Then the rv $Y = e^X$ is said to have a ***log-normal distribution with parameters*** μ ***and*** σ^2. Find the pdf of the log-normal.

SOLUTION:
$y = g(x) = e^x$ is a strictly increasing function of x and, since $\mathrm{supp}(f_X) = (-\infty, \infty)$, $\mathrm{supp}(f_Y) = (0, \infty)$.
$g^{-1}(y) = \ln(y) \implies \frac{d}{dy}(g^{-1}(y)) = \frac{1}{y} > 0$ when $y > 0$.
So, for $y > 0$,

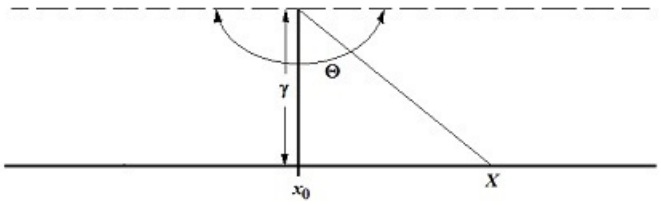
$$f_Y(y) = f_X(\ln(y)) \cdot \frac{1}{y} = \frac{e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}}}{y\sqrt{2\pi\sigma^2}}.$$

The pdf of the log-normal with parameters μ and σ^2 is, therefore,

$$f_Y(y) = \begin{cases} \dfrac{e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}}}{y\sqrt{2\pi\sigma^2}} & \text{for } y > 0 \\ 0 & \text{for } y \leq 0 \end{cases}.$$

Example. (the Cauchy distribution)
Suppose $\Theta \sim \mathrm{uniform}(-\frac{\pi}{2}, \frac{\pi}{2})$ and fix $x_0 \in \mathbb{R}$ and $\gamma > 0$.
Find the pdf of $X = x_0 + \gamma \tan(\Theta)$.

SOLUTION:



$f_\Theta(\theta) = \frac{1}{\pi}$ for $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$.
As θ ranges from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$, x ranges from $-\infty$ to ∞. Moreover,
$x = g(\theta) = x_0 + \gamma \tan(\theta) \implies \theta = g^{-1}(x) = \tan^{-1}\left(\frac{x-x_0}{\gamma}\right) \in (-\frac{\pi}{2}, \frac{\pi}{2})$ when $x \in \mathbb{R} \implies$
$\frac{d}{dx} g^{-1}(x) = \frac{1}{1+\left(\frac{x-x_0}{\gamma}\right)^2} \cdot \frac{1}{\gamma}$.

Therefore, for $-\infty < x < \infty$,

$$f_X(x) = \underbrace{f_\Theta\left(\tan^{-1}\left(\frac{x-x_0}{\gamma}\right)\right)}_{=\frac{1}{\pi}} \cdot \left(\frac{1}{1+\left(\frac{x-x_0}{\gamma}\right)^2}\right) \cdot \frac{1}{\gamma} = \frac{1}{\pi\gamma\left(1+\left(\frac{x-x_0}{\gamma}\right)^2\right)},$$

is the ***Cauchy distribution with parameters*** x_0 ***and*** γ abbreviated $X \sim \mathrm{Cauchy}(x_0, \gamma)$.
When $x_0 = 0$ and $\gamma = 1$, we call the $\mathrm{Cauchy}(0, 1)$ the ***standard Cauchy distribution***.

V. Jointly distributed random variables.

Jointly distributed random variables.

A natural extension to univariate probability, i.e., analysis of a single rv on Ω, is to multivariate probability, i.e., *many* rvs on Ω. When a collection of random variables are all defined on the *same sample space* we say the random variables are ***jointly distributed*** and that there is a sample space Ω that supports the joint collection.

Some examples of jointly distributed rvs.
• If the experiment is to roll a fair 6-sided die repeatedly and we define rvs X_i to be the value on the ith roll ($i = 1, 2, 3, \dots$), then this is a jointly distributed collection; in fact, since all these random variables are discrete we will say they are ***jointly discrete***.

• An experiment might select an individual within a certain population and, for this individual, we might measure their serum creatine level X, their heart rate Y, and their blood pressure Z. In this case the collection X, Y, Z would be jointly distributed; in fact, ***jointly continuous*** if these are all continuous rvs.

• An experiment might be to throw a dart at a dartboard n times. X (resp., Y) returns the x-coordinate (resp., y-coordinate) of where the dart lands on the dartboard. N might be the rv that counts the number of times the dart "misses" the dartboard completely. The collection X, Y, N are jointly distributed, but because these rvs are of mixed type (N is discrete whereas X and Y are continuous) we cannot say they are jointly discrete nor can we say they are jointly continuous, so we'll just say they are jointly distributed.

Jointly discrete rvs.

When we have a collection of discrete rvs all defined on the same sample space Ω we say these rvs are jointly discrete. To keep notation under control I will present the theory in the case of only *two* jointly discrete rvs, say X and Y.

If X and Y are jointly discrete, we define the ***joint probability mass function*** or, simply, the ***joint pmf*** by

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

The event $(X = x, Y = y)$ is shorthand for $\{\omega \in \Omega : X(\omega) = x\} \cap \{\omega \in \Omega : Y(\omega) = y\}$.

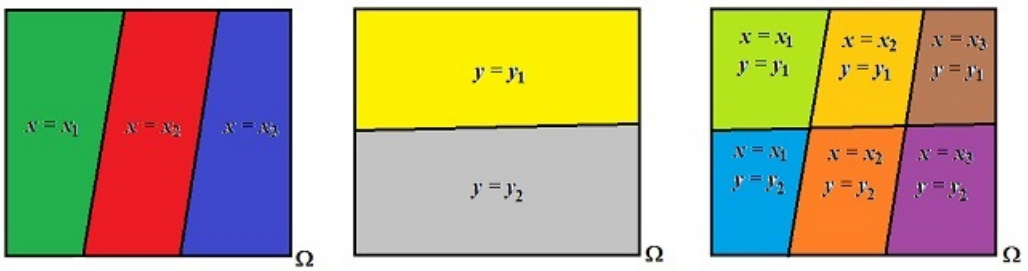


Figure. Each discrete rv X and Y partitions Ω, so do their intersections.

Example and discussion.
Suppose we have a box with 12 balls of which 3 are red, 4 are white, 5 are blue. We uniformly at random select 2 balls. Let X be the number of red balls, Y the number of blue balls. Construct the joint pmf of X and Y.

SOLUTION:
In this problem we seem to be only interested in the numbers of red and blue balls, but when we are drawing balls from this box it's fairly clear that we can draw red, blue and white balls. Since we are only drawing $n = 2$ balls and there are 3 red balls, X can take the values $0, 1, 2$; similarly, there are 5 blue balls, so Y can take the values $0, 1, 2$ as well:

$$X \in \{0, 1, 2\}, \quad Y \in \{0, 1, 2\}.$$

There are 12 balls total and selecting $n = 2$ uniformly at random means all $|\Omega| = \binom{12}{2}$ sample points are equally likely – I choose to ignore the order in which the balls are selected. The event $(X = x, Y = y)$ means that in our sample of size 2 we drew x red balls *and* y blue balls, *and*, therefore, $2 - x - y$ balls are neither red nor blue (i.e., white); thus, when ignoring the order in which the balls are selected, every collection of 2 balls can be decomposed in the union of 3 mutually exclusive events: the subset of red marbles drawn, the subset of blue marbles drawn, and the subset of marbles that are neither red nor blue – we allow some of these subsets to be empty, of course.

$$P(X = x, Y = y) = \frac{\binom{3}{x}\binom{5}{y}\binom{4}{2-x-y}}{\binom{12}{2}}.$$

I'll mention something that is hopefully obvious: in this problem $x + y$ needs to be less than of equal to 2 since we are only drawing 2 balls; and, in the case where $x + y > 2$ we necessarily have $P(X = x, Y = y) = 0$. Another way to have seen this is that when $x + y > 2$ the last binomial coefficient in the numerator is $\binom{4}{2-(x+y)}$ which would say we are selecting a *negative* number of objects from 4 objects and there must be 0 ways to do this!

Therefore, we get the following joint pmf:

$p_{X,Y}(x, y)$	$y = 0$	$y = 1$	$y = 2$
$x = 0$	$\frac{\binom{3}{0}\binom{5}{0}\binom{4}{2}}{\binom{12}{2}} = \frac{6}{66}$	$\frac{\binom{3}{0}\binom{5}{1}\binom{4}{1}}{\binom{12}{2}} = \frac{20}{66}$	$\frac{\binom{3}{0}\binom{5}{2}\binom{4}{0}}{\binom{12}{2}} = \frac{10}{66}$
$x = 1$	$\frac{\binom{3}{1}\binom{5}{0}\binom{4}{1}}{\binom{12}{2}} = \frac{12}{66}$	$\frac{\binom{3}{1}\binom{5}{1}\binom{4}{2}}{\binom{12}{0}} = \frac{15}{66}$	0
$x = 2$	$\frac{\binom{3}{2}\binom{5}{0}\binom{4}{0}}{\binom{12}{2}} = \frac{3}{66}$	0	0

Remark.
By the way, the probability distribution that we just constructed is an example of the ***multivariate hypergeometric distribution***. We will develop this distribution in a little more detail on page 184.

Computing probabilities involving jointly discrete rvs.
Analogous to the univariate discrete rv case, once we have the joint pmf computing probabilities involves summing the probability masses at each point (x, y) that belongs to the event.

Example (continued).
Following up with the example on the last page...
Now that we found the joint pmf of X and Y, let's compute
(a) $P(X \leq 1, Y \leq 1)$
(b) $P(X + Y \leq 1)$
(c) $P(X \leq 1)$
(d) $P(Y \leq 1)$

SOLUTION:
(a) The event $(X \leq 1, Y \leq 1)$ is $(x, y) = (0, 0), (0, 1), (1, 0), (1, 1)$; therefore, we just need to add the probability masses at these 4 points:

$$\frac{6}{66} + \frac{20}{66} + \frac{12}{66} + \frac{15}{66} = \frac{53}{66}.$$

(b) The event $(X + Y \leq 1)$ is $(x, y) = (0, 0), (0, 1), (1, 0)$; therefore, we just need to add the probability masses at these 3 points:

$$\frac{6}{66} + \frac{20}{66} + \frac{12}{66} = \frac{38}{66}.$$

(c) The event $(X \leq 1)$ is $(x, y) = (0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2)$; now there are 6 probability mass to add:

$$\frac{6}{66} + \frac{20}{66} + \frac{10}{66} + \frac{12}{66} + \frac{15}{66} + 0 = \frac{63}{66}.$$

(d) The event $(Y \leq 1)$ is $(x, y) = (0, 0), (0, 1), (1, 0), (1, 1), (2, 0), (2, 1)$; and, again, 6 (different) probability masses to add:

$$\frac{6}{66} + \frac{20}{66} + \frac{12}{66} + \frac{15}{66} + \frac{3}{66} + 0 = \frac{56}{66}.$$

Remark.
The joint pmf of a collection of rvs completely describes the probabilistic behavior (distribution) of the collection as a whole. What's important to know is that we can recover the distribution of any subcollection of these rvs (i.e, its ***marginal distribution***) from the joint pmf of the original collection.

Example (continued still).
Still following up with the last example...
Use the joint pmf to find the marginal pmf of X and the marginal pmf of Y.

SOLUTIONS:
For $x = 0, 1, 2$, $P(X = x) = \sum_y P(X = x, Y = y) = p_{X,Y}(x, 0) + p_{X,Y}(x, 1) + p_{X,Y}(x, 2)$:

x	0	1	2
$p_X(x)$	$\frac{36}{66}$	$\frac{27}{66}$	$\frac{3}{66}$

For $y = 0, 1, 2$, $P(Y = y) = \sum_x P(X = x, Y = y) = p_{X,Y}(0, y) + p_{X,Y}(1, y) + p_{X,Y}(2, y)$:

x	0	1	2
$p_X(x)$	$\frac{21}{66}$	$\frac{35}{66}$	$\frac{10}{66}$

Finding the marginal pmf.
Basically what we did to find the marginal pmf of a subset of variables is we went back into the original joint pmf and, for each fixed choice of the values for the variables in the subset we simply "summed out" the *other* variables not in this subset. Below, in red ink we have the marginal pmf of X, in blue ink we have the marginal pmf of Y:

$p_{X,Y}(x, y)$	$y = 0$	$y = 1$	$y = 2$	
$x = 0$	$\frac{\binom{3}{0}\binom{5}{0}\binom{4}{2}}{\binom{12}{2}} = \frac{6}{66}$	$\frac{\binom{3}{0}\binom{5}{1}\binom{4}{1}}{\binom{12}{2}} = \frac{20}{66}$	$\frac{\binom{3}{0}\binom{5}{2}\binom{4}{0}}{\binom{12}{2}} = \frac{10}{66}$	$\frac{36}{66} = P(X = 0)$
$x = 1$	$\frac{\binom{3}{1}\binom{5}{0}\binom{4}{1}}{\binom{12}{2}} = \frac{12}{66}$	$\frac{\binom{3}{1}\binom{5}{1}\binom{4}{2}}{\binom{12}{0}} = \frac{15}{66}$	0	$\frac{27}{66} = P(X = 1)$
$x = 2$	$\frac{\binom{3}{2}\binom{5}{0}\binom{4}{0}}{\binom{12}{2}} = \frac{3}{66}$	0	0	$\frac{3}{66} = P(X = 2)$
	$\frac{21}{66} = P(Y = 0)$	$\frac{35}{66} = P(Y = 1)$	$\frac{10}{66} = P(Y = 2)$	

Exercise for the student.
We have a fair coin. In each round, three people $1, 2, 3$ in this order take turns flipping the coin, and, after person 3 has flipped, the coin goes back to person 1 and we enter the next round; and, this repeats *ad infinitum*. We define the random variable $X = i$ if person i is the first to flip a head. We define Y to be the round on which a person flips the first head. Construct the joint pmf of X, Y.
Then, from this joint pmf, find the marginal pmfs for each of X and Y.

ANSWER:

$P(X=x, Y=y)$	$y=1$	$y=2$	$y=3$	$\cdots$	$y=n$	$\cdots$	$P(X=x)$
$x=1$	$\frac{1}{2}$	$\left(\frac{1}{2}\right)^4$	$\left(\frac{1}{2}\right)^7$	$\cdots$	$\left(\frac{1}{2}\right)^{1+3(n-1)}$	$\cdots$	$\frac{4}{7}$
$x=2$	$\left(\frac{1}{2}\right)^2$	$\left(\frac{1}{2}\right)^5$	$\left(\frac{1}{2}\right)^8$	$\cdots$	$\left(\frac{1}{2}\right)^{2+3(n-1)}$	$\cdots$	$\frac{2}{7}$
$x=3$	$\left(\frac{1}{2}\right)^3$	$\left(\frac{1}{2}\right)^6$	$\left(\frac{1}{2}\right)^9$	$\cdots$	$\left(\frac{1}{2}\right)^{3n}$	$\cdots$	$\frac{1}{7}$
$P(Y=y)$	$\frac{7}{8}$	$\frac{7}{8}\left(\frac{1}{8}\right)$	$\frac{7}{8}\left(\frac{1}{8}\right)^2$	$\cdots$	$\frac{7}{8}\left(\frac{1}{8}\right)^{n-1}$	$\cdots$	

What's really interesting in this example is that, unlike the previous example,

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad \text{for } \textit{all} \text{ choices of } x \text{ and } y.$$

We will learn that this means the random variables X and Y are ***independent***: knowledge of the value of one of these rvs will not influence the probability of the other. To compare you should note that in the red/blue balls example, our joint pmf doesn't have this property, and, therefore, those rvs are *dependent*.

Exercise for the student.
Show that the following function is a joint pmf, namely, when we sum all the probability masses we get 1:

$$p_{X,Y}(x, y) = \frac{2y}{x(x+1)}\left(\frac{1}{2}\right)^x \quad \text{for } x = 1, 2, 3, \ldots \ ; \ y = 1, 2, \ldots, x.$$

The formula $\sum_{y=1}^{x} y = \frac{x(x+1)}{2}$ might be helpful. Can you prove this?

Concluding remarks.
Thus far, I only presented examples involving *two* jointly discrete rvs; this was primarily to simplify presentation and notation. But, I would like to now present two common models of joint pmfs that involve more than two rvs typically. The first is the multivariate hypergeometric distribution and the other is the multinomial distribution.

Multivariate hypergeometric distribution.
As the name suggests this distribution is an extension of the hypergeometric distribution to $k > 2$ random variables. The situation is this:

We have a *finite* population of N distinct objects. We suppose these objects are comprised of k types. There are N_i objects of type i ($i = 1, 2, \ldots, k$), where $N_1 + N_2 + \cdots + N_k = N$. We sample uniformly at random n from the population *without replacement* and let X_i denote the number of type i's in our sample.
Then, if $n_1 + n_2 + \cdots + n_k = n$,

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_k}{n_k}}{\binom{N}{n}}.$$

Facts regarding the multivariate hypergeometric.
When $k = 2$, the above definition reduces to the ordinary hypergeometric – the $k = 2$ types would arbitrarily be called *successes* and *failures*. The univariate marginal distributions of a multivariate hypergeomtric are ordinary hypergeometric distributions.

Example.
A bin has 1000 components of which 750 are superior quality, 200 are above average quality, 40 are good quality, and 10 are poor quality. We uniformly at random sample 8 components. Let X_1 be the number of superior, X_2 the number of above average, X_3 the number of good, and X_4 the number of poor quality components in the sample. Compute the probability we get 2 of each quality classification. How about 6 superior, 2 above average, 0 good and 0 poor?

SOLUTION:

$$P(X_1 = 2, X_2 = 2, X_3 = 3, X_4 = 2) = \frac{\binom{750}{2}\binom{200}{2}\binom{40}{2}\binom{10}{2}}{\binom{1000}{8}} \approx 0.00000813551.$$

$$P(X_1 = 6, X_2 = 2, X_3 = 0, X_4 = 0) = \frac{\binom{750}{6}\binom{200}{2}\binom{40}{0}\binom{10}{0}}{\binom{1000}{8}} \approx 0.19993657.$$

Multinomial distribution.
As the name suggests this is an extension of the binomial distribution to $k > 2$ types. As an experiment leading to this distribution we can reconsider the multivariate hypergeometric experiment only this time we sample *with* replacement. Alternatively, we can think of our population as *infinite* but having proportions p_i of type i in it ($i = 1, 2, \ldots, k$), where $p_1 + p_2 + \cdots + p_k = 1$.

We randomly (draw independently one at a time) sample n from the population, and let X_i be the number of type i in our sample. If $n_1 + n_2 + \cdots + n_k = n$, then

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}.$$

Facts regarding the multinomial distribution.
When $k = 2$ this is the binomial distribution as long as we interepret only one of the two rvs as counting the number of "successes" – the other would be counting the number of "failures". The bivariate marginal pmfs of the multinomial distribution, i.e., the marginal pmfs involving any two rvs amongst $X_1, X_2, \ldots, X_k$, are binomial.
The multinomial theorem:

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_{n_1} \sum_{n_2} \cdots \sum_{n_k} \frac{n!}{n_1! n_2! \cdots n_k!} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k},$$

where the sum extends over all nonnegative integers $n_1, n_2, \ldots, n_k$ that sum to n.

Remark.
Much of what we learned regarding the similarities between the ordinary hypergeometric and binomial distributions goes over to these multivariate versions.

Example.
We have a weighted 6-sided die (not fair). On any roll, face i comes up with probability $p_i = \frac{i}{21}$ (notice that $p_1 + p_2 + \cdots + p_6 = 1$ here). We roll this die 210 times. Compute the probability that the number of times we see each face follows these proportions directly, i.e., $X_i = 10i$ for $i = 1, 2, \ldots, 6$.

SOLUTION:

$$P(X_1 = 10, X_2 = 20, X_3 = 30, X_4 = 40, X_5 = 50, X_6 = 60)$$
$$= \frac{210!}{10!20!30!40!50!60!} \left(\tfrac{1}{21}\right)^{10} \left(\tfrac{2}{21}\right)^{20} \left(\tfrac{3}{21}\right)^{30} \left(\tfrac{4}{21}\right)^{40} \left(\tfrac{5}{21}\right)^{50} \left(\tfrac{6}{21}\right)^{60} \approx 0.000005349.$$

Jointly continuous rvs
This section is a natural extension to what we did with univariate continuous rvs. There, we defined a continuous rv as one with continuous CDF on $\mathbb{R}$. We'll do the same here.

A collection of rvs $X_1, X_2, \ldots, X_n$ is said to be ***jointly continuous*** provided they are all defined on the same sample space *and* the joint CDF:

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

is a continuous function on $\mathbb{R}^n$.

Existence and characterization of the joint pdf:
Just as we did with the assumption on page 138, we will assume throughout this course that the ***joint probability density function*** or, simply, the ***joint pdf:***

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$$

exists for (almost) all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. Under this assumption, the resulting joint pdf can be characterized by these two properties:

1. $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ is *defined* on all of $\mathbb{R}^n$ *and* is *nonnegative*; and
2. $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n = 1.$

When we have such a collection of jointly continuous rvs its joint pdf is either given to us or is modeled by the specific situation at hand – like the collection of rvs are independent or we can model the conditional behavior of some of these rvs given the others, etc. More on this later!

Example and discussion.
Consider the function[7]

$$f(x,y) = \begin{cases} \frac{1}{2} & \text{for } 0 \le x \le 2, 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases} .$$

(a) Show this is a joint pdf.
(b) Use it to compute $P(X \le 1, Y \le \frac{1}{2})$.
(c) Compute $P(X + Y \le 1)$
(d) Compute $P(Y \le X^2)$.

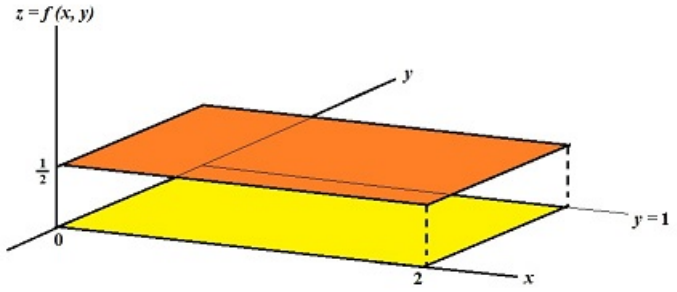


Figure. Plot of the pdf

SOLUTION:
(a) To show a function of two variables is a joint pdf, we need to show two things: first, that it's defined and nonnegative everywhere in $\mathbb{R}^2$, and, second, show that the integral of the function over $\mathbb{R}^2$ is 1.

The function $f(x,y)$ given is defined for all $(x,y) \in \mathbb{R}^2$ and, since $f(x,y)$ only takes on two values, namely, 0 and $\frac{1}{2}$, and both of these are nonnegative, $f(x,y) \ge 0$ everywhere. We now need to show

$$\int\int f(x,y)\, dA = 1.$$

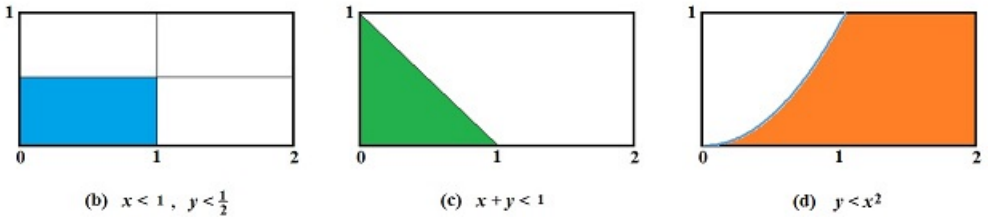
But from the plot of $f(x,y)$ we see a very simple geometry, the integral in this case is just the volume of the box under the orange surface over the yellow region: $2 \times 1 \times \frac{1}{2} = 1$.

Alternatively, we could have used calculus

$$\int\int f(x,y)\, dA = \int_0^1 \int_0^2 \frac{1}{2}\, dxdy = \frac{1}{2}\int_0^1 \int_0^2 dxdy = \frac{1}{2}\,\text{area}([0,2]\times[0,1]) = 1.$$

[7]this is the pdf of a uniform distribution on the rectangle $[0,2] \times [0,1]$ – see page 208

As nice as the plot of the surface $z = f(x, y)$ is, it's practically *useless*! Definitely not needed. What will be more important is a picture of the support of $f(x, y)$ – in this case, the rectangle $[0, 2] \times [0, 1]$:



We imagine that over each of the shaded regions in the above picture there is our surface $f(x, y)$, under which we are trying to compute the content, i.e., "volume" – this content is our probability.

In this example, $f(x, y) = \frac{1}{2}$ is a constant, so for any $\mathcal{R} \subseteq [0, 2] \times [0, 1]$, the integral of this function will be equal to $\frac{1}{2} \times \text{area}(\mathcal{R})$. If $\mathcal{R}$ has a geometric shape whose area is easily computed then we are essentially done. This is the case, for instance, parts (b) and (c):

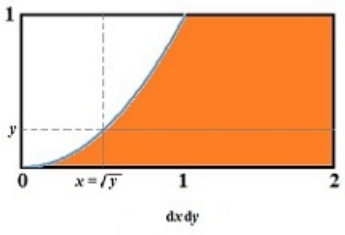
(b) $P(X \leq 1, Y \leq \frac{1}{2}) = \frac{1}{2} \times \text{area(blue box)} = \frac{1}{2} \times (1 \times \frac{1}{2}) = \frac{1}{4}$.

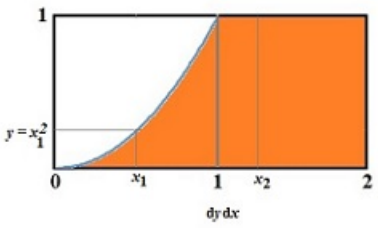
(c) $P(X + Y \leq 1) = \frac{1}{2} \times \text{area(green triangle)} = \frac{1}{2} \times (\frac{1}{2} \times 1 \times 1) = \frac{1}{4}$.

(d) For this part we don't have a well-recognized geometric shape, so we're going to have to resort to some calculus, i.e., we're going to need to integrate. We need to compute

$$\int \int_{\mathcal{R}} \frac{1}{2} \, dxdy.$$

Finding probability is, therefore, reduced to a calculus problem. We are integrating the function $f(x, y) = \frac{1}{2}$ over the orange region. We have a choice of parametrization of this integral: we can integrate, say, $dxdy$ or $dydx$.





If we parameterize $dxdy$, we get the integral:

$$\int_0^1 \int_{\sqrt{y}}^2 \frac{1}{2} \, dxdy = \frac{1}{2} \int_0^1 2 - y^{1/2} \, dy = \frac{1}{2} \left(2y - \frac{2}{3} y^{3/2} \Big|_{y=0}^{y=1} \right) = \frac{2}{3}.$$

If we parameterize $dydx$, we get the integral:

$$\int_0^1 \int_0^{x_1^2} \frac{1}{2} \, dydx_1 + \int_1^2 \int_0^1 \frac{1}{2} \, dydx_2 = \frac{1}{2} \int_0^1 x_1^2 \, dx_1 + \frac{1}{2} \int_1^2 1 \, dx_2 = \frac{1}{6} + \frac{1}{2} = \frac{2}{3}.$$

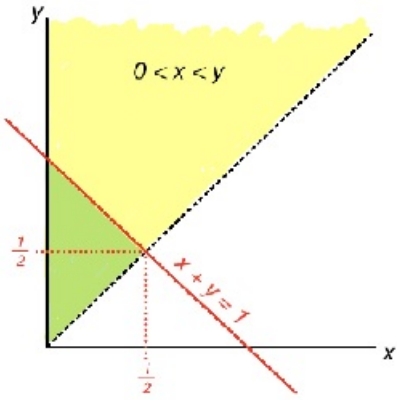
The reason the $dxdy$ integration was "nicer" in the last example is because with this parametrization the *entrance boundary* and the *exit boundary* are each described by single functions throughout the integration; whereas in the $dydx$ parametrization the *exit boundary* changes it function definition depending on whether $x < 1$ or $x > 1$: when $x < 1$ we enter the orange region through the line $y = 0$ and exit the region at $y = x^2$, but when $x > 1$ we enter the region, again, at $y = 0$ but exit the region at $y = 1$.

Example.
Suppose X and Y are jointly continuous rvs with the joint pdf

$$f_{X,Y}(x,y) = \begin{cases} e^{-y} & \text{for } 0 < x < y < \infty \\ 0 & \text{elsewhere} \end{cases} .$$

(Exercise: Show this is a pdf.) Compute
(a) $P(X + Y \leq 1)$
(b) $P(\frac{Y}{X} \leq u)$ for $u > 1$.

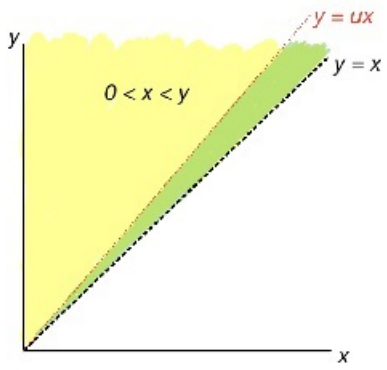
SOLUTION:
(a) Let's first sketch the support of the joint pdf together with the region of interest:



The green shaded region is the region of interest, i.e., we need to integrate the joint pdf over this region. We have a couple of choices of parametrization of this region, $dx\,dy$ or $dy\,dx$. By inspection of the green region the parametrization $dy\,dx$ will lead to a single integral whereas the parametrization $dx\,dy$ will lead to two integrals – this can be seen because once we fix y in the interval $0 < y < 1$, where we *exit* the green region in the x-direction will depend on whether $y < \frac{1}{2}$ or $y > \frac{1}{2}$. So, I choose to integrate $dy\,dx$ for this problem:

$$\begin{aligned} P(X + Y \leq 1) &= \int_0^{\frac{1}{2}} \int_x^{1-x} e^{-y}\, dy\, dx \\ &= \int_0^{\frac{1}{2}} -e^{-y}\Big|_{y=x}^{y=1-x}\, dx = \int_0^{\frac{1}{2}} e^{-x} - e^{-1+x}\, dx \\ &= -e^{-x} - e^{-1+x}\Big|_{x=0}^{x=1} = 1 - 2e^{-1/2} + e^{-1} \approx .1548. \end{aligned}$$

(b) Here's a sketch of the region where $y/x \leq u$:



In this case either parametrization seems fine. As an integral using $dxdy$:

$$P\left(\frac{Y}{X} \leq u\right) = \int_0^\infty \int_{y/u}^{y} e^{-y}\, dxdy$$

and as an integral using $dydx$:

$$P\left(\frac{Y}{X} \leq u\right) = \int_0^\infty \int_{x}^{ux} e^{-y}\, dydx.$$

Please verify that either of these integrals gives the probability $1 - \frac{1}{u}$. In fact, we've shown that the CDF of $\frac{Y}{X}$, namely, $F_{\frac{Y}{X}}(u) = P(\frac{Y}{X} \leq u) = 1 - \frac{1}{u}$ for $u > 1$. When $u \leq 1$, $F_{\frac{Y}{X}}(u) = 0$.

Marginal pdf.

If X and Y are jointly continuous with joint pdf $f_{X,Y}$, then we define the **marginal pdf** $f_X(x)$ **of** X by

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

and the **marginal pdf** $f_Y(y)$ **of** Y by

$$f_Y(y) := \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx.$$

Notice, for instance, to find the marginal pdf of X at the value x we "integrate out" the other variable for this fixed value of x. Similarly, the marginal pdf of Y at the value y is found by "integrating out" the joint pdf at this fixed value of y.

Just as with the marginal pmfs for jointly discrete rvs, the marginal pdfs represent the pdf of the of rv in the absence of the other variable.

Example.
Continuing with the example on page 189, we have

$$f_{X,Y}(x,y) = \begin{cases} e^{-y} & \text{for } 0 < x < y < \infty \\ 0 & \text{elsewhere} \end{cases} .$$

Find the marginal pdf of X and the marginal pdf of Y.

SOLUTION:
Let's first find the marginal pdf of X. So, *fix* a value x.

Since the support of the joint pdf is the set $0 < x < y < \infty$, if we pick a value of $x \leq 0$, the joint pdf is 0 and, therefore, $f_X(x) = 0$ when $x \leq 0$.

If $x > 0$, the we need to integrate out y for this fixed value of x. But the support requires $0 < x < y < \infty$, i.e., y cannot be smaller than x. So, when $x > 0$,

$$f_X(x) = \int_x^\infty e^{-y}\, dy = e^{-x}.$$

Therefore,

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

i.e., $X \sim \text{Exp}(1)$.

Now for the marginal of Y. *Fix* a value of y. Again, if this value of y is less than or equal to 0, the joint pdf will vanish and, in this case, $f_Y(y) = 0$.

On the other hand, if $y > 0$, the support of the joint pdf requires $0 < x < y$ so that when we integrate out the x variable we just need to integrate the joint for x between 0 and y, i.e., when $y > 0$

$$f_Y(y) = \int_0^y e^{-y}\, dx = ye^{-y}.$$

Thus, our marginal pdf of Y is

$$f_Y(y) = \begin{cases} ye^{-y} & \text{for } y > 0 \\ 0 & \text{for } y \leq 0 \end{cases},$$

i.e., $Y \sim \text{Gamma}(2, 1)$.

Remark.
When we have *more than just two* jointly continuous rvs, the idea of the marginal pdfs can be extended naturally. For instance, suppose we have five jointly continuous rvs X_1, X_2, X_3, X_4, X_5 with joint pdf $f_{X_1,X_2,X_3,X_4,X_5}(x_1, x_2, x_3, x_4, x_5)$. Then to find the marginal pdf of *any subset* of these five rvs we would "integrate out" all the other rvs not in the subset. For example,

$$f_{X_3}(x_3) = \int_{-\infty}^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty f_{X_1,X_2,X_3,X_4,X_5}(x_1, x_2, x_3, x_4, x_5)\, dx_1 dx_2 dx_4 dx_5$$

and

$$f_{X_2,X_4}(x_2, x_4) = \int_{-\infty}^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty f_{X_1,X_2,X_3,X_4,X_5}(x_1, x_2, x_3, x_4, x_5)\, dx_1 dx_3 dx_5,$$

etc. Analogies apply to jointly discrete rvs as well.

Jointly distributed rvs of mixed type.
There are many problems involving jointly distributed rvs that are of mixed type (say, discrete and continuous). The following example explores this situation.

Example.
Suppose we flip a fair coin repeatedly. Let X be the trial of the first head we obtain (so that $X \sim$ geometric($\frac{1}{2}$) - a discrete rv). Now, given $X = x$, suppose we win Y dollars, where $Y \sim \text{Exp}(x)$. In this scenario, the distribution of the amount of money we win is a *continuous* rv and Y depends on when we flip our first head, i.e., X.

Fix $h > 0$.

$$\begin{aligned}\frac{P(X = x, y - h < Y \leq y)}{h} &= P(X = x) \cdot \frac{P(y - h < Y \leq y | X = x)}{h} \\ &= \left(\frac{1}{2}\right)^x \cdot \frac{\int_{y-h}^{y} xe^{-xu}\, du}{h}\end{aligned}$$

and as $h \to 0$ the above converges to the function

$$p_{X,Y}(x, y) = \left(\frac{1}{2}\right)^x xe^{-xy} \quad \text{for } x = 1, 2, 3, \ldots;\ y > 0.$$

The function $p_{X,Y}(x, y)$ is neither a joint pmf nor a joint pdf: in fact, $p_{X,Y}(x, y)$ is a pmf in the x argument and is a pdf in the y argument. We will loosely call such a function the joint distribution of X and Y.

In this example, we were given the marginal pmf of X, in fact, X is geometric($\frac{1}{2}$). However, we were only given the *conditional* distribution of Y when the value of X is specified. We now ask: what is the marginal pdf of Y?

Since we know the joint distribution of X and Y, the marginal of Y can be found by "summing out" the X variable: Let $y > 0$. Then

$$\begin{aligned}f_Y(y) &= \sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^x xe^{-xy} \\ &= \left(\frac{e^{-y}}{2}\right) \sum_{x=1}^{\infty} x \left(\frac{e^{-y}}{2}\right)^{x-1} \\ &= \left(\frac{e^{-y}}{2}\right) \left(1 - \frac{e^{-y}}{2}\right)^{-1} \underbrace{\sum_{x=1}^{\infty} x \left(1 - \frac{e^{-y}}{2}\right) \left(\frac{e^{-y}}{2}\right)^{x-1}}_{=\text{mean of a geom}(1-\frac{e^{-y}}{2})=(1-\frac{e^{-y}}{2})^{-1}} \\ &= \left(\frac{e^{-y}}{2}\right) \left(1 - \frac{e^{-y}}{2}\right)^{-2}.\end{aligned}$$

Independence of random variables.

A collection of random variables is called **independent** if and only if the joint distribution of the collection factors into the product of its marginals.

For instance, jointly discrete rvs $X_1, X_2, \ldots, X_n$ are independent means the joint pmf of $X_1, X_2, \ldots, X_n$ factors in the the product of the marginal pmfs of each rv in this collection:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n)$$

for *every* choice of $x_1, x_2, \ldots, x_n$. The "for every choice" is important here because we want the functions of $x_1, x_2, \ldots, x_n$ to be the same!

Similarly, jointly continuous rvs $X_1, X_2, \ldots, X_n$ are independent means the joint pdf of $X_1, X_2, \ldots, X_n$ is the product of the marginal pdfs of each X_i in this collection:

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n) \quad \text{for every choice } x_1, x_2, \ldots, x_n.$$

Random variables that are not independent are said to be **dependent**.

Example.
From the example on page 191

$$f_{X,Y}(x, y) = \begin{cases} e^{-y} & \text{for } 0 < x < y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

and we found the marginal pdfs to be

$$f_X(x) = \begin{cases} e^{-x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

and

$$f_Y(y) = \begin{cases} ye^{-y} & \text{for } y > 0 \\ 0 & \text{for } y \leq 0 \end{cases}.$$

Notice that

$$f_X(x)f_Y(y) = \begin{cases} e^{-x}ye^{-y} & \text{for } x > 0,\ y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

is positive when $x > y$ whereas $f_{X,Y}(x, y) = 0$ when $x > y$. So, these functions are not the same and the rvs are dependent!

Remark.
When jointly discrete rvs X and Y are independent it would follow that

$$P(X = x, Y = y) = P(X = x)P(Y = y) \qquad \text{for } \textit{all } x \text{ and } y.$$

For example, look at the following joint pmf below given of *independent* rvs X and Y:

$p_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	
$x=0$	.03	.05	.02	$.1 = P(X=0)$
$x=1$	.06	.10	.04	$.2 = P(X=1)$
$x=2$	.09	.15	.06	$.3 = P(X=2)$
$x=3$	.12	.20	.08	$.4 = P(X=3)$
	$.3 = P(Y=0)$	$.5 = P(Y=1)$	$.2 = P(Y=2)$	

When X and Y are independent the joint pmf for each fixed x is proportional to the marginal pmf of Y. Likewise, the joint pmf for each fixed y is proportional to the marginal pmf of X. That is,

$$\frac{P(X = x, Y = y)}{P(X = x)} = P(Y = y|X = x) = P(Y = y)$$

and

$$\frac{P(X = x, Y = y)}{P(Y = y)} = P(X = x|Y = y) = P(X = x).$$

Example.
Suppose a rectangle has random edges lengths X and Y. Assume $X \sim \text{uniform}(0,1)$ and $Y \sim \text{uniform}(0,1)$ are independent. Compute the probability that the area of the rectangle is greater than $\frac{1}{2}$.

SOLUTION:
We are told that

$$f_X(x) = \begin{cases} 1 & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

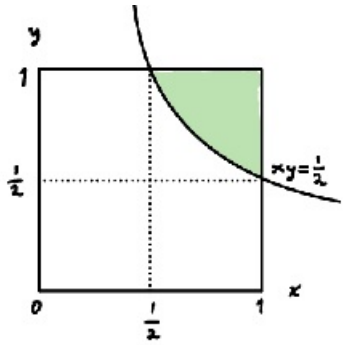
and

$$f_Y(y) = \begin{cases} 1 & \text{for } 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

and, therefore, the joint pdf of X and Y is (by assumed independence)

$$f_{X,Y}(x,y) = \begin{cases} 1 & \text{for } 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

We are interested in computing $P(XY > \frac{1}{2})$. Here's a sketch of the support and the event of interest:



$$\begin{aligned} P\left(XY > \frac{1}{2}\right) &= \int_{\frac{1}{2}}^{1} \int_{\frac{1}{2x}}^{1} 1 \, dy dx \\ &= \int_{\frac{1}{2}}^{1} 1 - \frac{1}{2x} \, dx \\ &= x - \frac{1}{2}\ln(x)\Big|_{x=\frac{1}{2}}^{1} \\ &= 1 - 0 - (\frac{1}{2} - \frac{1}{2}\ln(\frac{1}{2})) = \frac{1}{2} - \frac{1}{2}\ln(2). \end{aligned}$$

Exercise for the student.
Continue with the above example and show that the pdf of the area $A = XY$ of the rectangle is given by $f_A(a) = \begin{cases} -\ln(a) & \text{for } 0 < a < 1 \\ 0 & \text{elsewhere} \end{cases}$.

Convolution.

Mathematically, if we have two real-valued functions $f, g : \mathbb{R} \to \mathbb{R}$, the **convolution $f * g$ of f and g** is defined by

$$(f * g)(x) := \int_{-\infty}^{\infty} f(u)g(x-u)\,du. \tag{6}$$

Technically, (6) is called the **convolution integral of f and g**. We will soon show that when f and g are the pdfs of the *independent* continuous rvs X and Y, respectively, then $f * g$ is the pdf of $X + Y$.

There is a discrete analog to the convolution integral but it requires the discrete rvs to be *integer-valued*. Suppose X and Y are *independent* integer-valued rvs with respective pmfs $p_X(i) = P(X = i)$ and $p_Y(j) = P(Y = j)$. Then we will show the pmf of $X + Y$ is given by

$$\begin{aligned}(p_X * p_Y)(k) &:= \sum_{n=-\infty}^{\infty} p_X(n)p_Y(k-n) \\ &= \cdots + p_X(-1)p_Y(k+1) + p_X(0)p_Y(k) + p_X(1)p_Y(k-1) + \cdots\end{aligned} \tag{7}$$

Equation (7) is called the **discrete convolution of p_X and p_Y**.

The discrete convolution.

Let's start with the discrete case first: If X and Y are integer-valued, then $X + Y$ will be integer-valued. By the law of total probability

$$\begin{aligned}P(X+Y=k) &= \sum_{n=-\infty}^{\infty} P(X=n, X+Y=k) \\ &= \sum_{n=-\infty}^{\infty} P(X=n, n+Y=k) \\ &= \sum_{n=-\infty}^{\infty} P(X=n, Y=k-n) \\ &= \sum_{n=-\infty}^{\infty} \underbrace{P(X=n)}_{=p_X(n)} \underbrace{P(Y=k-n)}_{=p_Y(k-n)} =: (p_X * p_Y)(k).\end{aligned}$$

Exercise.
Show $(p_X * p_Y)(k) = (p_Y * p_X)(k)$ for every k, and, therefore, $p_X * p_Y$ can be defined as

$$(p_X * p_Y)(k) = \sum_{u=-\infty}^{\infty} P(X = k - u)P(Y = u).$$

Hint: When we employed the law of total probability to compute $P(X + Y = k)$ above we partitioned Ω by the events $(X = n)$ for integer n; try partitioning Ω by the events $(Y = u)$ for integer u instead. This exercise says that convolution is a commutative operation. The result shouldn't be shocking since the distribution of $X + Y$ should be the same as the distribution of $Y + X$.

Remark.
The convolution gives us a strategy for computing the distribution of a sum of independent random variables. Although the general definition of the discrete convolution has the sum extending over all integer indices from $-\infty$ to ∞, depending on the supports of the random variables involved for some indices the summand may vanish. Let's consider some common scenarios.

Scenario 1: X and Y have pmfs supported on the nonnegative integers. In this case $p_X(n) = 0$ if $n < 0$ and $p_Y(k - n) = 0$ when $k - n < 0$, i.e., when $n > k$. Therefore, for integer $k \geq 0$,

$$(p_X * p_Y)(k) = \sum_{n=-\infty}^{\infty} p_X(n)p_Y(k - n) = \sum_{n=0}^{k} p_X(n)p_Y(k - n).$$

Scenario 2: X is supported on the integers $x \geq m_1$ for some integer m_1, and the pmf of Y is supported on the integers $y \geq m_2$ for some integer m_2. In this case $p_X(n) = 0$ if $n < m_1$ and $p_Y(k - n) = 0$ when $k - n < m_2$, i.e., when $n > k - m_2$. Therefore, for integer $k \geq m_1 + m_2$,

$$(p_X * p_Y)(k) = \sum_{n=-\infty}^{\infty} p_X(n)p_Y(k - n) = \sum_{n=m_1}^{k-m_2} p_X(n)p_Y(k - n).$$

Scenario 3: X has nonnegative support but Y is supported on all integers. Then $p_X(n) = 0$ for $n < 0$. So, for $k \in \mathbb{Z}$,

$$(p_X * p_Y)(k) = \sum_{n=-\infty}^{\infty} p_X(n)p_Y(k - n) = \sum_{n=0}^{\infty} p_X(n)p_Y(k - n).$$

Example.(the neg.binom(r,p))
Prove that the sum of r independent geometric(p) rvs has a neg.binom(r,p) distribution.

SOLUTION:
Suppose $X_1 \sim$ geometric(p) and $X_2 \sim$ geometric(p) are independent. Recall

$$p(i) = P(X_1 = i) = P(X_2 = i) = p(1-p)^{i-1} \quad \text{for } i = 1, 2, 3, \ldots.$$

Let's find the distribution of their sum $X_1 + X_2$ as the discrete convolution. Since the support of the common distribution are integers $\geq m = 1$ we fall in scenario 2. In this case, for $k \geq 2$, we have

$$\begin{aligned} P(X_1 + X_2 = k) &= \sum_{n=1}^{k-1} P(X_1 = n) P(X_2 = k - n) \\ &= \sum_{n=1}^{k-1} p(1-p)^{n-1} \cdot p(1-p)^{k-n-1} \\ &= \sum_{n=1}^{k-1} p^2(1-p)^{k-2} = (k-1)p^2(1-p)^{k-2} \sim \text{neg.binom}(2,p). \end{aligned}$$

Continuing inductively, suppose that for some integer $r \geq 2$, whenever $X_1, X_2, \ldots, X_r$ are independent geometric(p) rvs,

$$X_1 + X_2 + \cdots + X_r \sim \text{neg.binom}(r,p).$$

Let $X_1, X_2, \ldots, X_r, X_{r+1}$ be independent geometric(p) rvs. Now, for integer $k \geq r+1$,

$$\begin{aligned} P(X_1 + \cdots + X_r + X_{r+1} = k) &= \sum_{n=r}^{k-1} P(X_1 + \cdots + X_r = n) P(X_{r+1} = k - n) \\ &= \sum_{n=r}^{k-1} \binom{n-1}{r-1} p^r (1-p)^{n-r} \cdot p(1-p)^{k-n-1} \\ &= \sum_{n=r}^{k-1} \binom{n-1}{r-1} p^{r+1} (1-p)^{k-(r+1)} \\ &= \binom{k-1}{r} p^{r+1} (1-p)^{k-(r+1)} \sim \text{neg.binom}(r+1,p), \end{aligned}$$

where the last equality follows from Fermat's combinatorial identity. Therefore, the sum of r independent geometric(p) rvs has a neg.binom(r,p) distribution.

Exercise for the student.
Suppose $X \sim$ Poisson(a) and $Y \sim$ Poisson(b) are independent. Both pmfs are both supported on the nonnegative integers. Use the discrete convolution to find the distribution of the sum $X + Y$.

Example.
$X \sim$ Bernoulli(p) and $Y \sim$ Bernoulli(q) are independent. Find the pmf of $X + Y$.

SOLUTION:
This can be done easily since the joint pmf of X and Y is easily constructed:

	$\mathbf{y = 0}$	$\mathbf{y = 1}$
$\mathbf{x = 0}$	$(1-p)(1-q)$	$(1-p)q$
$\mathbf{x = 1}$	$p(1-q)$	pq

The support of $X + Y$ is $\{0, 1, 2\}$, and
$P(X + Y = 0) = (1-p)(1-q)$,
$P(X + Y = 1) = p(1-q) + q(1-p)$,
$P(X + Y = 2) = pq$.

Exercise.
Show that discrete convolution is associative: $(p_X * p_Y) * p_Z = p_X * (p_Y * p_Z)$.

Example.
If $X \sim$ binom(n, p) and $Y \sim$ binom(m, p) are independent, show that $X + Y \sim$ binom$(n + m, p)$.

SOLUTION:
The support of $X + Y$ is $\{0, 1, 2, \ldots, n + m\}$. Fix a u in this support.

$$\begin{aligned}
P(X+Y=u) &= \sum_{k=-\infty}^{\infty} P(X=k)P(Y=u-k) \\
&= \sum_{k=0}^{u} \binom{n}{k} p^k (1-p)^{n-k} \cdot \binom{m}{u-k} p^{u-k} (1-p)^{m-u+k} \\
&= p^u (1-p)^{n+m-u} \underbrace{\sum_{k=0}^{u} \binom{n}{k}\binom{m}{u-k}}_{=\binom{n+m}{u} \text{ by normalization}} \\
&= \binom{n+m}{u} p^u (1-p)^{n+m-u},
\end{aligned}$$

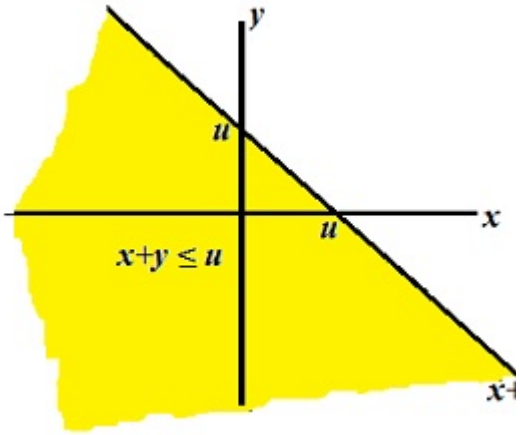
which shows $X + Y$ has the pmf of a binom$(n + m, p)$.

The convolution integral.

Let X and Y be continuous rvs – X having pdf $f(x)$ and Y having pdf $g(y)$. We will assume that X and Y are independent so that their joint pdf is $f(x)g(y)$. We can use the CDF method to find the pdf of $X+Y$. We start by writing down the CDF of $X+Y$:

$$F_{X+Y}(u) = P(X+Y \le u).$$

To compute this probability we are guided by the picture below:



$$\begin{aligned} F_{X+Y}(u) &= P(X+Y \le u) \\ &= \int_{-\infty}^{\infty} \Big(\underbrace{\int_{-\infty}^{u-x} f(x)g(y)\,dy}_{\text{a function of } u \text{ and } x} \Big) dx. \end{aligned}$$

Now we use the Liebniz rule

$$\begin{aligned} f_{X+Y}(u) &= \int_{-\infty}^{\infty} \frac{\partial}{\partial u} \int_{-\infty}^{u-x} f(x)g(y)\,dy\,dx \\ &= \int_{-\infty}^{\infty} f(x)g(u-x)\,dx =: (f*g)(u), \end{aligned}$$

and we see the pdf of $X+Y$ is the convolution integral of the individual pdfs.

Remark.

Just as in an earlier remark regarding the different scenarios in discrete convolutions, we may be able to restrict the integration from $-\infty$ to ∞ based on the support of the densities involved. For example, if X and Y have the same pdf f whose support is $[0,\infty)$, then $f(x) = 0$ for $x < 0$ and $f(u-x) = 0$ for $u - x < 0$, i.e., $u > x$, and consequently, the convolution reduces to

$$(f*f)(u) = \int_0^u f(x)f(u-x)\,dx$$

in this case.

Example.

Suppose X and Y are independent $\mathrm{N}(0,\sigma^2)$. Recall the pdf is $f(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$ for $-\infty < x < \infty$. Find the pdf of $X+Y$ using convolutions.

SOLUTION: The support of $X+Y$ is the entire real line, so let $u \in \mathbb{R}$.

$$\begin{aligned}
f_{X+Y}(u) &= \int_{-\infty}^{\infty} f(x)f(u-x)\,dx = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{e^{-\frac{(u-x)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\,dx \\
&= \frac{1}{2\pi\sigma^2}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{u^2+x^2-2ux}{2\sigma^2}}\,dx \\
&= \frac{1}{2\pi\sigma^2}\int_{-\infty}^{\infty} e^{-\frac{2x^2-2ux+u^2}{2\sigma^2}}\,dx \\
&= \frac{1}{2\pi\sigma^2}\int_{-\infty}^{\infty} e^{-\frac{2[x^2-ux+\frac{u^2}{4}-\frac{u^2}{4}+\frac{u^2}{2}]}{2\sigma^2}}\,dx \\
&= \frac{1}{2\pi\sigma^2}\int_{-\infty}^{\infty} e^{-\frac{2[(x-\frac{u}{2})^2+\frac{u^2}{4}]}{2\sigma^2}}\,dx \\
&= \frac{e^{-\frac{u^2}{4\sigma^2}}}{2\pi\sigma^2} \underbrace{\int_{-\infty}^{\infty} e^{-\frac{(x-\frac{u}{2})^2}{\sigma^2}}\,dx}_{=\sqrt{\pi\sigma^2} \text{ by normalization trick}} = \frac{e^{-\frac{u^2}{4\sigma^2}}}{\sqrt{4\pi\sigma^2}} \sim \mathrm{N}(0,2\sigma^2).
\end{aligned}$$

Example.
Suppose X and Y are independent $\mathrm{Exp}(\lambda)$ rvs. Find the pdf of $X+Y$ using convolutions.

SOLUTION: Since X and Y are supported on $[0,\infty)$ we can, for fixed $u > 0$, restrict the integration to $[0,u)$:

$$\begin{aligned}
f_{X+Y}(u) &= \int_{-\infty}^{\infty} f(x)f(u-x)\,dx \\
&= \int_0^u \lambda e^{-\lambda x}\lambda e^{-\lambda(u-x)}\,dx \\
&= \lambda^2 \int_0^u e^{-\lambda u}\,dx = \lambda^2 u e^{-\lambda u}.
\end{aligned}$$

So,

$$f_{X+Y}(u) = \begin{cases} \lambda^2 u e^{-\lambda u} & \text{for } u > 0 \\ 0 & \text{for } u \le 0 \end{cases},$$

which is the pdf of a Gamma$(2, \frac{1}{\lambda})$ distribution, i.e., an Erlang$(2,\lambda)$.

Exercise for the student.
Continue the last example to show inductively that if $X_1, X_2, \ldots X_n$ are independent Exp(λ) then

$$X_1 + X_2 + \cdots + X_n \sim \text{Gamma}(n, \frac{1}{\lambda}) = \text{Erlang}(n, \lambda).$$

The next example illustrates the nuances in computing the convolution.

Example.
Suppose $X \sim \text{unif}(0, 2)$ and $Y \sim \text{unif}(0, 3)$ are independent rvs. Use the convolution to find the pdf of $X + Y$.

SOLUTION: The support of $X + Y$ is $[0, 5]$, so let $0 < u < 5$.

$$\begin{aligned} f_X(x) &= \frac{1}{2}1_{(0,2)}(x) \quad \text{for } 0 < x < 2, \quad \text{and} \\ f_Y(y) &= \frac{1}{3}1_{(0,3)}(y) \quad \text{for } 0 < y < 3. \\ f_{X+Y}(u) &= \int_{-\infty}^{\infty} \frac{1}{2}1_{(0,2)}(x) \cdot \frac{1}{3}1_{(0,3)}(u - x)\, dx. \end{aligned}$$

The integrand is $\frac{1}{6}$ when both $0 < x < 2$ and $0 < u - x < 3$ ($u - 3 < x < u$). Therefore,

$$\max\{0, u - 3\} < x < \min\{2, u\}.$$

When $0 < u < 2$, we have $0 < x < u$, and $f_{X+Y}(u) = \int_0^u \frac{1}{6}\, dx = \frac{u}{6}$.
When $2 < u < 3$, we have $0 < x < 2$, and $f_{X+Y}(u) = \int_0^2 \frac{1}{6}\, dx = \frac{1}{3}$.
When $3 < u < 5$, we have $u - 3 < x < 2$, and $f_{X+Y}(u) = \int_{u-3}^2 \frac{1}{6}\, dx = \frac{5-u}{6}$.

$$f_{X+Y}(u) = \begin{cases} \frac{u}{6} & \text{for } 0 < u < 2 \\ \frac{1}{3} & \text{for } 2 < u < 3 \\ \frac{5-u}{6} & \text{for } 3 < u < 5 \\ 0 & \text{elsewhere} \end{cases}.$$

Here's a plot of this pdf:

Remark.
The idea of the convolution can still apply when the rvs are *not* independent. I will illustrate when X and Y are jointly continuous with joint pdf $f(x,y)$.

$$\begin{aligned} F_{X+Y}(u) &= P(X+Y \le u) \\ &= \int_{-\infty}^{\infty} \underbrace{\int_{-\infty}^{u-x} f(x,y)\,dy}_{\text{funtion of } u}\,dx. \end{aligned}$$

Now apply the Liebniz rule:

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f(x, u-x)\,dx.$$

Exercise for the student.
Suppose X and Y are jointly continuous with joint pdf $f_{X,Y}(x,y) = (x+y)1_{(0,1)}(x)1_{(0,1)}(y)$. Find the pdf of $X+Y$. Hint: since these rvs are *not* independent use the previous remark and when you compute the integral consult the example involving the uniform pdfs on page 202.

Important facts about independent rvs.
If X and Y are independent, then $h_1(X)$ and $h_2(Y)$ are independent for any functions h_1 and h_2. More generally, if

$$X_1, X_2, \ldots, X_k, X_{k+1}, \ldots, X_n$$

are independent, then

$$h_1(X_1, X_2, \ldots, X_k) \quad \text{and} \quad h_2(X_{k+1}, \ldots, X_n)$$

are also independent for any functions h_1 and h_2.

The law of the unconscious statistician naturally extends to functions of many jointly distributed random variables. We only present the result when we have just two jointly distributed rvs the analogous result holds for more than two rvs.

Law of the Unconscious Statistician re-visited.
Suppose X and Y are jointly distributed and $g : \mathbb{R}^2 \to \mathbb{R}$ is a real-valued function. Then

$$E[g(X,Y)] = \sum_{y=-\infty}^{\infty} \sum_{x=-\infty}^{\infty} g(x,y) P(X=x, Y=y) \qquad \text{if } X, Y \text{ are jointly discrete}$$

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\, dxdy \qquad \text{if } X, Y \text{ are jointly continuous}$$

assuming the expressions shown exist.

Example.
X, Y are jointly continuous with joint pdf $f_{X,Y}(x,y) = e^{-y}$ for $0 < x < y$. Find $E\left(\frac{X}{Y}\right)$.

SOLUTION:

$$\begin{aligned} E\left(\frac{X}{Y}\right) &= \int_0^{\infty} \int_0^{y} \frac{x}{y} e^{-y}\, dxdy \\ &= \int_0^{\infty} \frac{x^2}{2y} e^{-y} \Big|_{x=0}^{x=y} dy \\ &= \frac{1}{2} \int_0^{\infty} y e^{-y}\, dy = \frac{1}{2}. \end{aligned}$$

The following is a very useful **corollary:**
When X and Y are independent and the function $g(x,y) = h_1(x)h_2(y)$ is separable

$$E[h_1(X)h_2(Y)] = E[h_1(X)]E(h_2(Y)].$$

Proof: We'll prove this in the case where the rvs are independent and jointly continuous so that the joint pdf is $f_X(x)f_Y(y)$. The proof for the jointly discrete case is similar.

$$\begin{aligned} E[h_1(X)h_2(Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_1(x)h_2(y) f_X(x) f_Y(y)\, dxdy \\ &= \int_{-\infty}^{\infty} h_2(y) f_Y(y) \underbrace{\int_{-\infty}^{\infty} h_1(x) f_X(x)\, dx}_{=E[h_1(X)]}\, dy \\ &= E[h_1(X)] \int_{-\infty}^{\infty} h_2(y) f_Y(y) dy = E[h_1(X)]E[h_2(Y)]. \end{aligned}$$

□

An important and immediate application of this corollary applies to moment generating functions which we discuss now.

Moment generating functions - part 3.

Recall that if the MGF of an rv exists it is given by $M_X(\theta) = E(e^{\theta X})$.

Theorem.
Let X and Y be jointly distributed rvs that each possess an MGF. Then

$$M_{X+Y}(\theta) = M_X(\theta)M_Y(\theta).$$

Proof:

$$\begin{aligned} M_{X+Y}(\theta) &= E(e^{\theta(X+Y)}) \\ &= E(e^{\theta X}e^{\theta Y}) \\ &= E(e^{\theta X})E(e^{\theta Y}) \quad \text{(previous corollary applied to } h_1(x) = e^{\theta x},\ h_2(y) = e^{\theta y}) \\ &= M_X(\theta)M_Y(\theta). \end{aligned}$$

□

Remark.
The last theorem is often used in conjunction with the following important property of moment generating functions:

> *If U and V have moment generating functions that agree and are finite in an open neighborhood of $\theta = 0$, then U and V have the same probability distribution (or law).*

We may use the notation $U \stackrel{d}{=} V$ or $U \stackrel{\mathcal{L}}{=} V$ to mean that U and V have the same distribution/law,

Here are two examples we did using convolutions:

Example.
Suppose $X \sim \text{binom}(n, p)$ and $Y \sim \text{binom}(m, p)$ are independent. Find the distribution of $X + Y$.

SOLUTION:
We know

$$M_X(\theta) = (1 - p + pe^{\theta})^n \quad \text{and} \quad M_Y(\theta) = (1 - p + pe^{\theta})^m.$$

By the theorem,

$$\begin{aligned} M_{X+Y}(\theta) &= M_X(\theta)M_Y(\theta) \\ &= (1 - p + pe^{\theta})^n(1 - p + pe^{\theta})^m = (1 - p + pe^{\theta})^{n+m}, \end{aligned}$$

which is the moment generating function for a $\text{binom}(n + m, p)$ distribution. Therefore, $X + Y$ must have a $\text{binom}(n + m, p)$ distribution.

Example.
Suppose X_1, X_2 are independent Exp(λ) rvs. Find the distribution of $X_1 + X_2$.

SOLUTION:
We know

$$M_{X_1}(\theta) = M_{X_2}(\theta) = \left(1 - \frac{\theta}{\lambda}\right)^{-1}.$$

By the theorem,

$$\begin{aligned} M_{X+Y}(\theta) &= M_{X_1}(\theta) M_{X_2}(\theta) \\ &= \left(1 - \frac{\theta}{\lambda}\right)^{-1} \left(1 - \frac{\theta}{\lambda}\right)^{-1} = \left(1 - \frac{\theta}{\lambda}\right)^{-2}, \end{aligned}$$

which is the moment generating function for a Gamma($2, \frac{1}{\lambda}$) distribution. Therefore, $X_1 + X_2$ must have a Gamma($2, \frac{1}{\lambda}$) distribution.

Remark.
A closer look at the proof of the theorem will show that the theorem remains true for *any* finite number of (not just two) rvs: if $X_1, X_2, \ldots, X_n$ are independent Exp(λ)[8], then

$$M_{\sum_{i=1}^n X_i}(\theta) = \prod_{i=1}^n M_{X_i}(\theta) = \prod_{i=1}^n \left(1 - \frac{\theta}{\lambda}\right)^{-1} = \left(1 - \frac{\theta}{\lambda}\right)^{-n},$$

which is the moment generating function of a Gamma($n, \frac{1}{\lambda}$) and, therefore, $\sum_{i=1}^n X_i \sim$ Gamma($n, \frac{1}{\lambda}$).

Notation/terminology:
Very often we deal with random variables that are independent *and* all have the same probability distribution, and in this case we will say the random variables are **iid** which is short for *independent and identically distributed*.

[8] A good mnemonic: the sum of i.i.d. exponentials has a Gamma distribution.

Conditional distributions.

The general problem here is when we have a jointly distributed collection of rvs, we may want to find the distribution of some function of these given information about the values of a subcollection.

We first discuss the case of two jointly discrete rvs. We need to recall the *conditional probability formula* from page 73:

$$\textit{For any event } A\textit{, when } P(B) > 0\textit{, } P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

What makes this formula so useful is that the conditioning event involving discrete rvs are typically events of positive probability.

Example.
Suppose X and Y are independent with

$$X \sim \text{binom}(N_1, p) \quad \text{and} \quad Y \sim \text{binom}(N_2, p).$$

Find the conditional distribution of X given $X + Y = n$.

SOLUTION:
We are looking for $P(X = k|X + Y = n)$ as a function of k. We may use the notation

$$p_{X|X+Y}(k|n) := P(X = k|X + Y = n).$$

Since $X + Y = n$ we assume $n \in \text{supp}(X + Y) = \{0, 1, \ldots, N_1 + N_2\}$. It should be clear that $k = 0, 1, \ldots, n$. Also, from the examples involving the binomial on pages 199 and 205, we see the sum $X + Y$ has a binom$(N_1 + N_2, p)$ distribution. Now, applying the conditional probability formula above, we obtain

$$\begin{aligned}
P(X = k|X + Y = n) &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} \\
&= \frac{P(X = k, Y = n - k)}{\binom{N_1+N_2}{n} p^n (1-p)^{N_1+N_2-n}} \\
&= \frac{P(X = k)P(Y = n - k)}{\binom{N_1+N_2}{n} p^n (1-p)^{N_1+N_2-n}} \\
&= \frac{\binom{N_1}{k} p^k (1-p)^{N_1-k} \binom{N_2}{n-k} p^{n-k} (1-p)^{N_2-n+k}}{\binom{N_1+N_2}{n} p^n (1-p)^{N_1+N_2-n}} \\
&= \frac{\binom{N_1}{k}\binom{N_2}{n-k}}{\binom{N_1+N_2}{n}}
\end{aligned}$$

which is the hypergeometric distribution!

Remark.
In the last example, we are given that $X+Y=n$ occurred. If $P(X+Y=n)=0$ then, since $X+Y$ is discrete, the event $(X+Y=n)$ could not have occurred, and, therefore, it must be the case that $P(X+Y=n)>0$. This allowed us to use the conditional probability formula.

We now do an example involving jointly continuous rvs.

Example. (uniform distribution over the region $D \subseteq \mathbb{R}^2$)
Suppose X and Y are jointly continuous and that $(X,Y) \sim$ uniform(D) which means X, Y has the joint pdf

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\text{area}(D)} & \text{for } (x,y) \in D \\ 0 & \text{elsewhere} \end{cases}.$$

This is the model where every point $(x,y) \in D$ is equally likely. To fix ideas let's suppose that D is the unit disk in the plane centered at $(0,0)$: $D = \{(x,y) : x^2+y^2 \leq 1\}$ – see figure. Compute the probability that $Y \geq \frac{1}{2}$ given $X \geq \frac{1}{2}$.

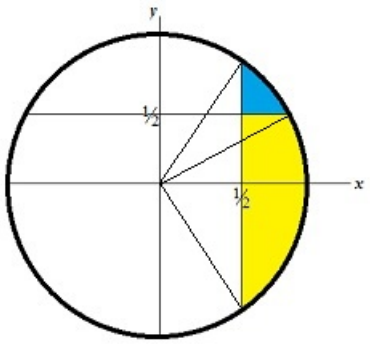


Figure. support of the uniform on the unit disk, blue region is $\{x \geq \frac{1}{2}\} \cap \{y \geq \frac{1}{2}\}$.

SOLUTION:
We are trying to compute $P(Y \geq \frac{1}{2} | X \geq \frac{1}{2})$. Since $P(X \geq \frac{1}{2}) > 0$ we apply the conditional probability formula:

$$P(Y \geq \tfrac{1}{2} | X \geq \tfrac{1}{2}) = \frac{P(X \geq \frac{1}{2}, Y \geq \frac{1}{2})}{P(X \geq \frac{1}{2})}.$$

Much like the example we did on page 187 when dealing with a uniform distribution of a region, probability is just the proportion of the region occupied by the event. From this perspective, the easiest way to evaluate these probabilities is to exploit the geometry of the situation – for example, compute areas of sectors and triangles and use similar triangles to get edge lengths.

I leave for the student to check that

$$P(X \geq \tfrac{1}{2}, Y \geq \tfrac{1}{2}) = \tfrac{1}{12} - \tfrac{\sqrt{3}-1}{4\pi} \text{ and } P(X \geq \tfrac{1}{2}) = \tfrac{1}{3} - \tfrac{\sqrt{3}}{4\pi},$$

and $P(Y \geq \frac{1}{2} | X \geq \frac{1}{2}) \approx .128$.

In the last example $P(Y \geq \frac{1}{2}|X \geq \frac{1}{2})$ was computed using the conditional probability formula as the conditioning event has positive probability. Let's continue with this example but try to compute

$$P(Y \geq \tfrac{1}{2}|X = \tfrac{1}{2})$$

instead. We run into a problem: we *cannot* use the conditional probability formula since $P(X = \frac{1}{2}) = 0$.

We now try to rectify this situation.

Conditional pdfs.

If X and Y have the joint pdf $f_{X,Y}(x, y)$ and marginal pdf $f_X(x)$, then for any x for which $f_X(x) > 0$ we define the ***conditional pdf*** $f_{Y|X}(y|x)$ ***of*** Y ***given*** $X = x$ by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

In this formula we think of the value of x as *fixed* and treat this conditional density as a function in the variable y.

How to use conditional densities to compute conditional probabilities having conditioning event $X = x$?

Fact:
If X, Y are jointly continuous with joint pdf $f_{X,Y}(x, y)$, then for any $a < b$

$$P(a < Y < b|X = x) = \int_a^b f_{Y|X}(y|x)\, dy.$$

Specifically, the answer to the above question is: we integrate the corresponding conditional pdf as a function of y fixing the value x.

Example. *Continuing with the last example...*
If we are interested in computing $P(Y \geq \frac{1}{2}|X = \frac{1}{2})$ then we recognize that the rv defining the conditioning event is X and the rv in the event of interest is Y and therefore, we'd need to find the conditional density of Y given $X = x$ with the value $x = \frac{1}{2}$.

We first compute the marginal $f_X(x)$: When $-1 < x < 1$ we have

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

Notice when $x = \frac{1}{2}$, we have $f_X(\frac{1}{2}) = \frac{\sqrt{3}}{\pi} > 0$. Therefore, for $-\frac{\sqrt{3}}{2} < y < \frac{\sqrt{3}}{2}$,

$$f_{Y|X}(y|\tfrac{1}{2}) = \frac{f_{X,Y}(\frac{1}{2}, y)}{f_X(\frac{1}{2})} = \frac{\frac{1}{\pi}}{\frac{\sqrt{3}}{\pi}} = \frac{1}{\sqrt{3}},$$

i.e., $Y|X = \frac{1}{2} \sim \text{uniform}(-\frac{\sqrt{3}}{2}, \frac{\sqrt{3}}{2})$.

Finally, to compute $P(Y \geq \frac{1}{2}|X = \frac{1}{2})$ we integrate $f_{Y|X}(y|\frac{1}{2})$ from $y = \frac{1}{2}$ to $y = \frac{\sqrt{3}}{2}$ (to stay in the support of this conditional density when $x = \frac{1}{2}$):

$$P(Y \geq \tfrac{1}{2}|X = \tfrac{1}{2}) = \int_{\frac{1}{2}}^{\frac{\sqrt{3}}{2}} \frac{1}{\sqrt{3}}\, dy = \frac{\sqrt{3}-1}{2\sqrt{3}}.$$

Remark.
When the conditioning event is a continuous rv equal to a single value we ***must*** first find the conditional distribution given this event and then use this conditional distribution to compute the conditional probability. On the other hand, if the conditioning event has positive probability, then we use the conditional probability formula to find the conditional probability. These two situations are *not* the same; for instance, in the last example

$$P(Y \geq \tfrac{1}{2}|X \geq \tfrac{1}{2}) = \frac{\int_{\frac{1}{2}}^{\frac{\sqrt{3}}{2}} \int_{\frac{1}{2}}^{\sqrt{1-x^2}} f_{X,Y}(x,y)\, dydx}{\int_{\frac{1}{2}}^{1} f_X(x)\, dx}$$

is not the same as

$$\int_{\frac{1}{2}}^{\frac{\sqrt{3}}{2}} P(Y \geq \tfrac{1}{2}|X = x)\, dx = \int_{\frac{1}{2}}^{\frac{\sqrt{3}}{2}} \int_{\frac{1}{2}}^{\sqrt{1-x^2}} f_{Y|X}(y|x)\, dydx.$$

The student should show that these two expressions do not evaluate to the same value. Generally speaking, a conditional pdf is only a pdf in the first variable and it is *not* a pdf in the conditioning variable.

Example.
Continuing from the example on page 191...
Suppose X and Y are jointly continuous with joint pdf $f(x, y) = e^{-y}$ for $0 < x < y < \infty$. Compute
(a) the conditional density of X given $Y = y$
(b) $P(X \geq 1|Y = 2)$
(c) $P(X \geq 1|Y \geq 2)$

SOLUTION:
(a) The marginal pdf of Y from page 191 is $f_Y(y) = ye^{-y}$ for $y > 0$ and, therefore,

$$f_{X|Y}(x|y) = \frac{e^{-y}}{ye^{-y}} = \frac{1}{y} \quad \text{for } 0 < x < y,$$

i.e., $X|Y = y \sim \text{uniform}(0, y)$.
(b) When $y = 2$, $X|Y = 2 \sim \text{uniform}(0, 2)$. Thus,

$$P(X \geq 1|Y = 2) = \int_1^\infty f_{X|Y}(x|2)\, dx = \int_1^2 \frac{1}{2}\, dx = \frac{1}{2}.$$

I leave as an **exercise** for the student to do part (c), the answer is $P(X \geq 1|Y \geq 2) = \frac{2}{3}$.

Since the conditional pdf of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

we can easily solve for the joint pdf of X and Y:

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y).$$

A common practice in probability and statistics (easily in Bayesian statistics) is to model the conditional distribution of one rv given another, and then specify the (marginal) distribution of the other like in this next example.

Example.
Suppose $X|Y = y \sim \text{Exp}(y)$ and $Y \sim \text{Gamma}(\alpha, 1)$.
(a) Write down the joint pdf of X and Y.
(b) From the joint pdf you found in part (a) compute the marginal pdf of X.

SOLUTION:
(a) We are told $f_{X|Y}(x|y) = ye^{-yx}$ for $x > 0$ and $f_Y(y) = \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}$ for $y > 0$. Therefore,

$$f_{X,Y}(x,y) = ye^{-yx} \cdot \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)} = \frac{y^{\alpha}e^{-(1+x)y}}{\Gamma(\alpha)} \quad \text{for } x > 0, y > 0.$$

(b) Fix an $x > 0$. Then

$$f_X(x) = \int_0^\infty \frac{y^{\alpha}e^{-(1+x)y}}{\Gamma(\alpha)}\, dy = \frac{1}{\Gamma(\alpha)} \int_0^\infty y^{\alpha}e^{-(1+x)y}\, dy.$$

We recognize the integrand of the last integral as being the functional form of a Gamma pdf having shape parameter $\alpha + 1$ and scale parameter $\frac{1}{x+1}$. By the normalization trick it follow that

$$\frac{1}{\Gamma(\alpha)} \int_0^\infty y^{\alpha}e^{-(1+x)y}\, dy = \frac{1}{\Gamma(\alpha)} \left(\frac{1}{x+1}\right)^{\alpha+1} \Gamma(\alpha+1) = \frac{\alpha}{(x+1)^{\alpha+1}}.$$

I just so happens that this pdf is recognizable pdf: X is Pareto(α).

The next example works with rvs of mixed types.

Example.
Suppose $X|\Theta \sim \text{binom}(n, \Theta)$ and $\Theta \sim \text{uniform}(0, 1)$. Derive
(a) the joint distribution of X, Θ,
(b) the marginal distribution of X,
(c) the conditional distribution of Θ given $X = x$.

SOLUTION (sketch):
We're told $f_\Theta(\theta) = 1$ for $0 < \theta < 1$ and $p_{X|\Theta}(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ for $x = 0, 1, \ldots, n$.
(a) $p_{X,\Theta}(x, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ for $x = 0, 1, \ldots, n,\ 0 < \theta < 1$.
(b) Fix $x \in \{0, 1, \ldots, n\}$.

$$\begin{aligned}
p_X(x) &= \int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}\,d\theta \\
&= \binom{n}{x}\int_0^1 \theta^x(1-\theta)^{n-x}\,d\theta \\
&= \binom{n}{x}\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}\underbrace{\int_0^1 \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^x(1-\theta)^{n-x}\,d\theta}_{=1 \text{ since it is Beta}(x+1, n-x+1)} \\
&= \binom{n}{x}\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} \\
&= \frac{n!}{x!(n-x)!}\frac{x!(n-x)!}{(n+1)!} = \frac{1}{n+1},
\end{aligned}$$

i.e., $X \sim$ discrete uniform on $\{0, 1, \ldots, n\}$.
(c) Assuming $x \in \{0, 1, \ldots, n\}$ is fixed,

$$f_{\Theta|X}(\theta|x) = \frac{p_{X,\Theta}(x,\theta)}{p_X(x)} = \frac{\frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}}{\frac{1}{n+1}} = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^x(1-\theta)^{n-x}$$

for $0 < \theta < 1$ which says $\Theta|X = x \sim \text{Beta}(x+1, n-x+1)$.

Exercise for the student.
Suppose $X|Y = y \sim \text{uniform}(-y, y)$ and $Y \sim \text{Gamma}(2, 1)$. Find the pdf of X.
Do you recognize this distribution?

Law of total probability – revisited.
Just as we learned earlier in the course, the law of total probability can be an effective tool to compute unconditional probabilities.

Example.
Let $X, Y, Z \sim$ iid Exp(1). Compute $P(X + Y < Z)$.

SOLUTION:
On one hand

$$P(X+Y<Z) = \int\int\int_{0<x+y<z} e^{-x}e^{-y}e^{-z}\,dxdydz$$

and we can try to parametrize the region. But, using the law of total probability we also have

$$\begin{aligned} P(X+Y<Z) &= \int_{-\infty}^{\infty} P(X+Y<Z|Z=z)f_Z(z)\,dz \\ &= \int_0^{\infty} P(X+Y<Z|Z=z)e^{-z}\,dz. \end{aligned}$$

Now, the term $P(X+Y<Z|Z=z) = P(X+Y<z|Z=z) = P(X+Y<z)$ since Z is independent of $X+Y$. Recalling that $X+Y \sim \text{Gamma}(2,1)$, i.e., has density $f(u) = ue^{-u}$ for $u>0$, we have

$$P(X+Y<z) = \int_0^z ue^{-u}\,du = 1-e^{-z}-ze^{-z} \quad \text{(details omitted)}.$$

We substitute this expression into the integral above

$$\begin{aligned} P(X+Y<Z) &= \int_0^{\infty} P(X+Y<z)e^{-z}\,dz \\ &= \int_0^{\infty} (1-e^{-z}-ze^{-z})e^{-z}\,dz \\ &= \int_0^{\infty} e^{-z}\,dz - \int_0^{\infty} e^{-2z}\,dz - \int_0^{\infty} ze^{-2z}\,dz \\ &= 1-\frac{1}{2}-\frac{1}{4} \\ &= \frac{1}{4}, \end{aligned}$$

where we used the normalization trick in the last integral.

We come back now to the problem of finding the distribution of a transformation of a continuous random variable or jointly continuous rvs (if there's more than one). We've already had some exposure to the CDF method, method of convolutions, and the moment generating function method. We now discuss another method, but initially present the 2-dimensional version.

Method of Jacobians.

Suppose we know the joint pdf $f_{X,Y}(x,y)$ of X and Y and we consider the rvs U and V defined by

$$U = g_1(X,Y) \quad \text{and} \quad V = g_2(X,Y).$$

What's the joint pdf $f_{U,V}(u,v)$ of U and V?

We assume the mapping $(x,y) \mapsto (u,v) := (g_1(x,y), g_2(x,y))$ is continuously differentiable and one-to-one (so that it is invertible). In this case it follows $x = h_1(u,v)$ and $y = h_2(u,v)$ and the answer to the question is given by the following

Theorem. (Jacobian transformation in 2-d)
Suppose X, Y are jointly continuous with joint pdf $f_{X,Y}$ having support $\mathcal{A} = \text{supp}(f_{X,Y})$ and

$$u = g_1(x,y) \quad \text{and} \quad v = g_2(x,y)$$

is a continuously differentiable one-to-one transformation of $\mathcal{A}$ into $\mathcal{B}$. Then, if we denote the inverse transformation by

$$x = h_1(u,v) \quad \text{and} \quad y = h_2(u,v),$$

the joint pdf of $U = g_1(X,Y)$ and $V = g_2(X,Y)$ is given by

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v)) \cdot |J|,$$

where

$$J := \det \begin{pmatrix} \frac{\partial h_1(u,v)}{\partial u} & \frac{\partial h_1(u,v)}{\partial v} \\ \frac{\partial h_2(u,v)}{\partial u} & \frac{\partial h_2(u,v)}{\partial v} \end{pmatrix} = \frac{\partial h_1(u,v)}{\partial u}\frac{\partial h_2(u,v)}{\partial v} - \frac{\partial h_1(u,v)}{\partial v}\frac{\partial h_2(u,v)}{\partial u}$$

is the ***Jacobian determinant***.

Remark.
The CDF method and method of Jacobians are the two most widely used methods for finding the distribution of functions of continuous random variables.

Example.
Suppose X, Y are independent Exp(1) rvs so that $f_{X,Y}(x,y) = e^{-x}e^{-y}$ for $x > 0,\ y > 0$. Derive the joint pdf of U, V, where

$$U = X + Y \quad \text{and} \quad V = Y.$$

SOLUTION:
The transformation $u = x + y$ and $v = y$ is invertible:

$$x = u - v =: h_1(u,v) \quad \text{and} \quad y = v =: h_2(u,v).$$

Since $x > 0$, $u - v > 0 \implies u > v$. Since $y > 0$, $v > 0$. Therefore, the support of U, V is $0 < v < u < \infty$. The Jacobian of the inverse transformation is

$$J = \det \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = 1 \implies |J| = 1.$$

Consequently, the joint pdf of U, V is

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(u-v, v)|J| \\ &= e^{-(u-v)}e^{-v} \cdot 1 \\ &= \begin{cases} e^{-u} & \text{for } 0 < v < u < \infty \\ 0 & \text{elsewhere} \end{cases} . \end{aligned}$$

Remark.
At this point we can compute the marginal pdf of U:

$$f_U(u) = \int_0^u e^{-u}\, dv = ue^{-u} \quad \text{for } u > 0.$$

What if the transformation had been $U = X + Y$ and $V = X - Y$ instead? Derive the joint pdf of U, V now.

I'll sketch a solution; I'll leave the details to the student.
The transformation $u = x + y,\ v = x - y$ has the inverse transformation $x = \frac{1}{2}(u+v)$ and $y = \frac{1}{2}(u-v)$. The support of U, V is $u > 0,\ -u < v < u$. The Jacobian is $J = -\frac{1}{2}$.

$$f_{U,V}(u,v) = e^{-\frac{1}{2}(u+v)}e^{-\frac{1}{2}(u-v)}|J| = \frac{1}{2}e^{-u} \quad \text{for } u > 0,\ -u < v < u.$$

Notice that the marginal pdf of U is the same as before (as we should). Also, the student can check:

$$f_V(v) = \int_{|v|}^{\infty} \frac{1}{2}e^{-u}\, du = \frac{1}{2}e^{-|v|} \quad \text{for } v > 0,$$

i.e., V has a double-exponential (Laplace) distribution.

As these last examples show, if the goal was to find a marginal distribution, then we can still use the method of Jacobians. This may require us to first find or create artificial variables to obtain a one-to-one transformation. It will not matter how we choose the artificial variable(s) as when we integrate them out we will arrive to the same marginal.

Example. (The Beta distribution)
Suppose $X_1 \sim \text{Gamma}(\alpha_1, 1)$ and $X_2 \sim \text{Gamma}(\alpha_2, 1)$ are *independent*.
Find the pdf of $U = \frac{X_1}{X_1+X_2}$.

SOLUTION:
To apply the method of Jacobians we would need to introduce another rv, say, V to make the transformation $(x_1, x_2) \mapsto (u, v)$ one-to-one. The choice of V will, in fact, not matter. So, for now we'll choose $V = X_1 + X_2$.

The joint pdf of X_1, X_2 is $f_{X_1,X_2}(x_1, x_2) = \frac{x_1^{\alpha_1-1}e^{-x_1}}{\Gamma(\alpha_1)} \cdot \frac{x_2^{\alpha_2-1}e^{-x_2}}{\Gamma(\alpha_2)}$ for $x_1 > 0, x_2 > 0$. The inverse transformation is $x_1 = uv$, $x_2 = v - uv = (1-u)v$ with Jacobian

$$J = \det \begin{pmatrix} v & u \\ -v & 1-u \end{pmatrix} = v.$$

Since $x_1 > 0$ and $x_2 > 0$, we have $u > 0$, $v = x_1 + x_2 > 0$, and $(1-u)v > 0 \implies u < 1$. Therefore, the support of U, V is $0 < u < 1$ and $v > 0$. For such u, v,

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(uv, (1-u)v)|v| \\ &= \frac{[uv]^{\alpha_1-1}e^{-uv}}{\Gamma(\alpha_1)} \cdot \frac{[(1-u)v]^{\alpha_2-1}e^{-(1-u)v}}{\Gamma(\alpha_2)} \cdot v \quad (\text{since } |v| = v > 0) \\ &= \frac{u^{\alpha_1-1}v^{\alpha_1-1}e^{-uv}(1-u)^{\alpha_2-1}v^{\alpha_2-1}e^{-(1-u)v}v}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \\ &= \frac{u^{\alpha_1-1}(1-u)^{\alpha_2-1} \cdot v^{\alpha_1+\alpha_2-1}e^{-v}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \quad \text{for } 0 < u < 1,\ v > 0. \end{aligned}$$

The marginal pdf of U is

$$\begin{aligned} f_U(u) &= \int_0^\infty \frac{u^{\alpha_1-1}(1-u)^{\alpha_2-1} \cdot v^{\alpha_1+\alpha_2-1}e^{-v}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\, dv \\ &= \frac{u^{\alpha_1-1}(1-u)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \underbrace{\int_0^\infty v^{\alpha_1+\alpha_2-1}e^{-v}\, dv}_{=\Gamma(\alpha_1+\alpha_2)} \\ &= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} u^{\alpha_1-1}(1-u)^{\alpha_2-1} \quad \text{for } 0 < u < 1. \end{aligned}$$

That is, $U = \frac{X_1}{X_1+X_2} \sim \text{Beta}(\alpha_1, \alpha_2)$. In fact,

$$f_{U,V}(u,v) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} u^{\alpha_1-1}(1-u)^{\alpha_2-1} \cdot \frac{v^{\alpha_1+\alpha_2-1}e^{-v}}{\Gamma(\alpha_1+\alpha_2)}$$

and we see that $f_{U,V}(u,v) = f_U(u)f_V(v)$ with $V \sim \text{Gamma}(\alpha_1+\alpha_2, 1)$ and U and V are *independent*!

Exercise for the student.
Please re-do this example by instead choosing $V = X_1$. Show that you still get the same Beta marginal for U, however, U and V will *not* be independent in this case.

The method of Jacobians naturally extends from 1- and 2-dimensions to d-dimensions. We present the following theorem which may require the reader know a bit of linear algebra to compute the Jacobian determinant.

Theorem. (Jacobian transformation in finite dimensions)
Suppose $X_1, X_2, \ldots, X_d$ are jointly continuous with joint pdf $f = f_{X_1,\ldots,X_d}$ having support $\mathcal{A} = \text{supp}(f)$ and

$$u_1 = g_1(x_1, x_2, \ldots, x_d), \ldots, \ u_d = g_d(x_1, x_2, \ldots, x_d)$$

is a continuously differentiable one-to-one transformation of $\mathcal{A}$ into $\mathcal{B}$. Then, if we denote the inverse transformation by

$$x_1 = h_1(u_1, \ldots, u_d), \ \ldots, \ x_d = h_d(u_1, \ldots, u_d),$$

the joint pdf of $U_1, U_2, \ldots, U_d$ is given by

$$f_{U_1,\ldots,U_d}(u_1, \ldots, u_d) = f_{X_1,\ldots,X_d}(x_1, \ldots, x_d) \cdot |J|,$$

where

$$J := \det \begin{pmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \cdots & \frac{\partial x_1}{\partial u_d} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \cdots & \frac{\partial x_2}{\partial u_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial u_1} & \frac{\partial x_d}{\partial u_2} & \cdots & \frac{\partial x_d}{\partial u_d} \end{pmatrix}$$

is the ***Jacobian determinant***.

We now will generalize the last example that led us to the Beta distribution.

Example. (The Dirichlet distribution)
Let X_i, $i = 1, 2, \ldots, d+1$ be independent rvs and, for each i, suppose $X_i \sim \text{Gamma}(\alpha_i, 1)$. Define

$$U_i = \frac{X_i}{X_1 + X_2 + \cdots + X_{d+1}} \quad \text{for } i = 1, 2, \ldots, d, \quad \text{and} \quad U_{d+1} = X_1 + X_2 + \cdots + X_{d+1}.$$

Show that the d-variate marginal pdf of $U_1, \ldots, U_d$ is

$$f_{U_1,\ldots,U_d}(u_1, \ldots, u_d) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_{d+1})}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{d+1})} u_1^{\alpha_1 - 1} \cdots u_d^{\alpha_d - 1} (1 - u_1 - \cdots - u_d)^{\alpha_{d+1} - 1}$$

This is the so-called ***Dirichlet distribution***.

SOLUTION:
I'll leave the general case for the reader, I will instead do the case where $d + 1 = 3$. We have the transformation $u_1 = \frac{x_1}{x_1+x_2+x_3}$, $u_2 = \frac{x_2}{x_1+x_2+x_3}$, $u_3 = x_1 + x_2 + x_3$. The inverse transformation:

$$x_1 = u_1 u_3, \quad x_2 = u_2 u_3, \quad \text{and} \quad x_3 = u_3 - u_1 u_3 - u_2 u_3 = (1 - u_1 - u_2) u_3.$$

Since $x_1 > 0, x_2 > 0$, and $x_3 > 0$, $u_1 > 0, u_2 > 0, u_3 > 0$ and since $u_1 + u_2 = \frac{x_1+x_2}{x_1+x_2+x_3} < 1$ we have the support of U_1, U_2:

$$u_1 > 0, u_2 > 0, u_1 + u_2 < 1.$$

The Jacobian is

$$\begin{aligned} J &= \det \begin{pmatrix} u_3 & 0 & u_1 \\ 0 & u_3 & u_2 \\ -u_3 & -u_3 & 1-u_1-u_2 \end{pmatrix} \\ &= \det \begin{pmatrix} u_3 & 0 & u_1 \\ 0 & u_3 & u_2 \\ 0 & -u_3 & 1-u_2 \end{pmatrix} \\ &= \det \begin{pmatrix} u_3 & 0 & u_1 \\ 0 & u_3 & u_2 \\ 0 & 0 & 1 \end{pmatrix} \\ &= u_3^2. \end{aligned}$$

In this computation we used the linear algebra result that we do not change the determinant of a matrix by replacing a row by a multiple of another row added to it. We went from the first determinant to the next determinant by adding the first row to the last row. We went from the second determinant to the third determinant by adding the second row to the third row. The last determinant is a triangular matrix whose determinant is the product of the major diagonal entries.

Finally, the joint pdf of U_1, U_2, U_3 is

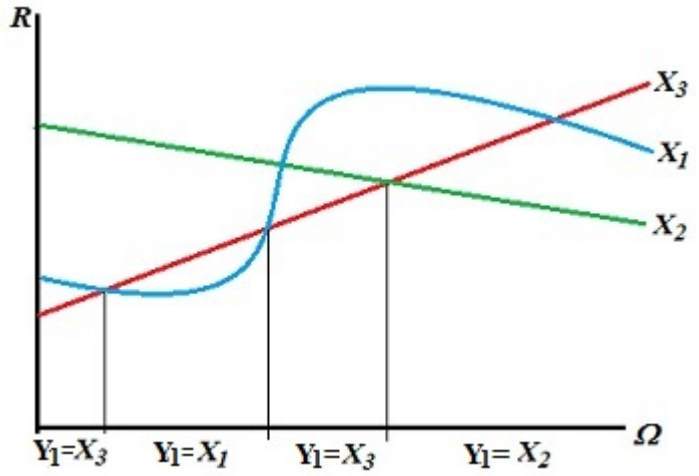
$$\begin{aligned} f(u_1,u_2,u_3) &= f_{X_1,X_2,X_3}(u_1u_3, u_2u_3, u_3)|u_3^2| \\ &= \frac{[u_1u_3]^{\alpha_1-1}e^{-u_1u_3}[u_2u_3]^{\alpha_2-1}e^{-u_2u_3}[(1-u_1-u_2)u_3]^{\alpha_3-1}e^{-(1-u_1-u_2)u_3}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \cdot u_3^2 \\ &= \frac{u_1^{\alpha_1-1}u_2^{\alpha_2-1}(1-u_1-u_2)^{\alpha_3-1}u_3^{\alpha_1+\alpha_2+\alpha_3-1}e^{-u_3}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \end{aligned}$$

and the (bivariate) marginal pdf of U_1, U_2 is

$$\begin{aligned} f_{U_1,U_2}(u_1,u_2) &= \int_0^\infty \frac{u_1^{\alpha_1-1}u_2^{\alpha_2-1}(1-u_1-u_2)^{\alpha_3-1}u_3^{\alpha_1+\alpha_2+\alpha_3-1}e^{-u_3}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}\,du_3 \\ &= \frac{u_1^{\alpha_1-1}u_2^{\alpha_2-1}(1-u_1-u_2)^{\alpha_3-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \underbrace{\int_0^\infty u_3^{\alpha_1+\alpha_2+\alpha_3-1}e^{-u_3}\,du_3}_{=\Gamma(\alpha_1+\alpha_2+\alpha_3)} \\ &= \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} u_1^{\alpha_1-1}u_2^{\alpha_2-1}(1-u_1-u_2)^{\alpha_3-1}. \end{aligned}$$

Ordered statistics.

Suppose we have a jointly distributed collection of random variables, say, $X_1, X_2, \ldots, X_n$. Since they are all defined on the same sample space Ω, we can define the "new" random variables Y_j to be the jth smallest among $X_1, X_2, \ldots, X_n$. The random variables Y_j are called the ***ordered statistics*** of the sample $X_1, X_2, \ldots, X_n$. The ordered statistics are well-defined: for each $\omega \in \Omega$, $X_i(\omega) \in \mathbb{R}$ for $i = 1, 2, \ldots, n$ and, therefore, for any $j = 1, 2, \ldots, n$, $Y_j(\omega)$ being the jth smallest value among them makes complete sense. It is important to note that the value of Y_j depends on the entire collection of X_i's and, therefore, Y_j is a function of all these rvs $X_1, X_2, \ldots, X_n$. Moreover, the Y_j's satisfy the condition $Y_1 \leq Y_2 \leq \cdots \leq Y_n$.



For instance, for the rv Y_1, the value $Y_1(\omega)$ is $\min\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}$ and this minimum value will come from that X_i which happens to be the smallest among $X_1, X_2, \ldots, X_n$ for the particular ω. The picture shows $n = 3$ rvs X_1, X_2 and X_3 as mappings from Ω into $\mathbb{R}$. Abstractly, from left to right, for those ω in the first section, X_3 is the smallest of the three rvs so $Y_1 = X_3$ in this region, followed next by X_1 being the smallest so that $Y_1 = X_1$ in this region, followed by $Y_1 = X_3$ again, followed by $Y_1 = X_2$.

Simplifying assumption we make when dealing with ordered statistics:
$X_1, X_2, \ldots, X_n$ are *independent and identically distributed* (iid) with CDF $F(x)$.

1. (univariate) distributions of the minimum Y_1 and the maximum Y_n.
The purpose of this section is to first analyze the distributions associated with the two *extreme* ordered statistics: the minimum and the maximum:

$$Y_1 = \min\{X_1, X_2, \ldots, X_n\} \quad \text{and} \quad Y_n = \max\{X_1, X_2, \ldots, X_n\}.$$

useful device when working with Y_1 and/or Y_n:
The minimum of a collection is greater than x if and only if *every* member of the collection is greater than x:

$$(Y_1 > x) = (X_1 > x, X_2 > x, \ldots, X_n > x),$$

and, the maximum of a collection is less than or equal to x if and only if *every* member of the collection is less than or equal to x:

$$(Y_n \leq x) = (X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x).$$

CDF of the minimum Y_1:

$$\begin{aligned}
F_{Y_1}(y) &= P(Y_1 \leq y) \\
&= 1 - P(Y_1 > y) \\
&= 1 - P(X_1 > y, X_2 > y, \ldots, X_n > y) \\
&= 1 - P(X_1 > y)P(X_2 > y)\cdots P(X_n > y) \quad \text{(using independence)} \\
&= 1 - \big(1 - P(X_1 \leq y)\big)\big(1 - P(X_2 \leq y)\big)\cdots\big(1 - P(X_n \leq y)\big) \\
&= 1 - \big(1 - F(y)\big)\big(1 - F(y)\big)\cdots\big(1 - F(y)\big) \quad \text{(identically distributed)} \\
&= 1 - (1 - F(y))^n.
\end{aligned}$$

CDF of the maximum Y_n:

$$\begin{aligned}
F_{Y_n}(y) &= P(Y_n \leq y) \\
&= P(X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y) \\
&= P(X_1 \leq y)P(X_2 \leq y)\cdots P(X_n \leq y) \quad \text{(using independence)} \\
&= P(X_1 \leq y)P(X_1 \leq y)\cdots P(X_1 \leq y) \quad \text{(identically distributed)} \\
&= F(y)^n.
\end{aligned}$$

pdfs of Y_1 and of Y_n when the iid collection is continuous
If the iid rvs $X_1, X_2, \ldots, X_n$ are absolutely continuous – so that $F(x)$ possesses a pdf $f(x)$ – then we can go further and say that

$$f_{Y_1}(y) = n\, f(y)[1 - F(y)]^{n-1}$$

$$f_{Y_n}(y) = n\, f(y)[F(y)]^{n-1}.$$

Advice:
It is often easier to re-derive the above calculations when trying to find the pdfs for Y_1 and Y_n than it is to remember these formulas.

Important Example. (The minimum of iid exponentials is, again, exponential)
Suppose $X_1, X_2, \ldots, X_n \sim$ iid Exp(λ). Show that $Y_1 \sim \text{Exp}(n\lambda)$.

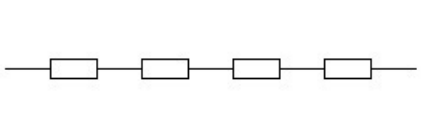
SOLUTION:
Here, $F(x) = 1 - e^{-\lambda x}$ for $x > 0$, and $f(x) = \lambda e^{-\lambda x}$ for $x > 0$. Therefore, for $y > 0$,

$$\begin{aligned} F_{Y_1}(y) &= 1 - P(Y_1 > y) \\ &= 1 - P(X_1 > y)P(X_2 > y)\cdots P(X_n > y) \\ &= 1 - (1 - F(y))^n \\ &= 1 - (e^{-\lambda y})^n = 1 - e^{-n\lambda y}, \end{aligned}$$

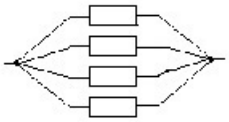
which shows Y_1 has the CDF of an Exp($n\lambda$). From here we can easily take the derivative to see

$$f_{Y_1}(y) = n\lambda e^{-n\lambda y} \quad \text{for } y > 0.$$

Application: components in series and in parallel.
Suppose a system is comprised of a sequence of n identical components in series (so when the first component fails the system fails). As an example suppose we have a system with $n = 4$ identical components hooked up in series and the lifetime of each component follows an exponential distribution with mean lifetime of 8 years (i.e., a rate $\lambda = 1/8$), also assume component failures are independent.



components arranged in series: first component to fail leads to system failure



components in parallel: last component to fail leads to system failure

Let X_1, X_2, X_3, X_4 represent the lifetimes of these 4 components. Then $Y_1 = \min\{X_1, X_2, X_3, X_4\}$ represents the time until the first failure. The last example showed us that $Y_1 \sim \text{Exp}(1/2)$. Moreover, the probability the system is functioning after 6 months ($= 1/2$ year) is

$$P(Y_1 > \tfrac{1}{2}) = \textstyle\int_{\frac{1}{2}}^{\infty} \frac{1}{2} e^{-y/2}\, dy = e^{-1/4} \approx .7788.$$

What if, instead of 4 of these components in series, we had only 2 of them in series? Then the distribution of Y_1 would now be Exp(1/4) and $P(Y_1 > \frac{1}{2})$ would now be $\approx .8825$.

Now suppose the 4 components are hooked up in parallel so that the system fails at the time Y_4, i.e., when the last component fail (and, therefore, all components have failed). Recall each X_i has CDF $F(x) = 1 - e^{-x/8}$ for $x > 0$ and, therefore, $F_{Y_4}(y) = (1 - e^{-y/8})^4$.

$$P(Y_4 > \tfrac{1}{2}) = 1 - F_{Y_4}(\tfrac{1}{2}) = 1 - (1 - e^{-1/16})^4 \approx .99987.$$

Exercise for the student.
Suppose $X_1, X_2, X_3, \ldots$ is a sequence of iid uniform$(0,1)$. Fix $x > 0$ and let $Y_1^{(n)} = \min\{X_1, X_2, \ldots, X_n\}$ be the *running minimum*. Find a formula for $P(Y_1^{(n)} > \frac{x}{n})$. You may assume $n > x$ so that $\frac{x}{n} < 1$. What happens to this expression as n tends to ∞? What can you say about the (limiting) distribution of $nY_1^{(n)} := n \min\{X_1, X_2, \ldots, X_n\}$ as $n \to \infty$, i.e., $F_{nY_1^{(n)}}(x) = P(nY_1^{(n)} \leq x)$ as $n \to \infty$? Does his CDF look familiar?

Example.
We have a balanced 1000-sided die with faces having distinct numbers from 1 thru 1000. We plan to roll it 500 times. For $i = 1, 2, \ldots, 500$, let X_i be the value showing on the ith roll. Compute the probability the largest value rolled is *greater than* 995, i.e., any one of the 5 values from $996, 997, 998, 999, 1000$? Estimate this probability using a limiting result we learned.

SOLUTION:
In this problem X_1, X_2, X_{500} are iid discrete uniform on the set $\{1, 2, \ldots, 1000\}$, and the largest value $Y_{500} = \max\{X_1, X_2, \ldots, X_{500}\}$. The probability of interest is

$$\begin{aligned} P(Y_{500} > 995) &= 1 - P(Y_{500} \leq 995) \\ &= 1 - (\frac{995}{1000})^{500}. \end{aligned}$$

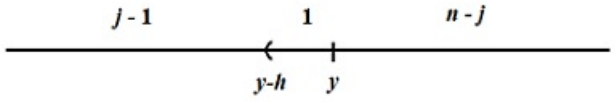
We can estimate this probability by invoking the limiting representation of e^u from page 106:

$$1 - (\frac{995}{1000})^{500} = 1 - \left(1 - \frac{5}{1000}\right)^{\frac{1000}{2}} = 1 - \left(\left(1 - \frac{5}{1000}\right)^{1000}\right)^{\frac{1}{2}} \approx 1 - \left(e^{-5}\right)^{\frac{1}{2}} \ (\approx .9179).$$

2. marginal distribution of the jth ordered statistic Y_j.
Y_j is the rv that takes the jth smallest value produced by $X_1, X_2, \ldots, X_n$. On the last few pages of these notes we discussed the cases $j = 1$ and $j = n$ in general, but we now will generalize to arbitrary j from 1 to n. Of course, we continue with the iid assumption, but first present the case where the rvs are continuous before handling the general CDF case.

2.1 case of iid continuous rvs.
Since the case where $X_1, X_2, \ldots, X_n$ are iid continuous rvs having CDF $F(x)$ and pdf $f(x)$ is of special important we first start with this case. This case allows for a "simpler" analysis of the ordered statistics since continuous rvs have a probability 0 of being equal. This means when the joint collection is observed we may assume their values are *distinct* (and thus separated by a positive distance).

Let $X_1, X_2, \ldots, X_n$ be iid continuous rvs having CDF $F(x)$ and pdf $f(x)$ and, for fixed $j = 1, 2, \ldots, n$, consider the jth ordered statistic Y_j. For fixed *small* $h > 0$ – going to be sent to 0 – the event $(y - h < Y_j \leq y)$ means the jth smallest among $X_1, X_2, \ldots, X_n$ lies in the interval $(y - h, y]$. Since the variables are continuous and (presumably) h is small and going to be sent to 0, the event $(y - h < Y_j \leq y)$ is equivalent to saying exactly *one* of the n variables $X_1, X_2, \ldots, X_n$ lies in the interval $(y - h, y]$, *and* $j - 1$ of the remaining variables lie in $(-\infty, y - h]$, *and* the remaining $n - j$ variables lie in (y, ∞). See the picture.



So, the event $(y - h < Y_j \leq y)$ is equivalent to

$$\Big\{\text{one of the } n \ X_i's \text{ in } (y-h, y]\Big\} \cap \Big\{j-1 \ X_i's \leq y - h\Big\} \cap \Big\{\text{remaining } X_i's > y\Big\}.$$

Therefore,

$$P(y - h < Y_j \leq y) \ = \ \binom{n}{j-1, 1, n-j} F(y-h)^{j-1} \big(F(y) - F(y-h)\big) \big(1 - F(y)\big)^{n-j}$$

and, consequently, if we divide both sides by h and send $h \to 0+$:

$$\begin{aligned} f_{Y_j}(y) &= \lim_{h \to 0+} P(y - h < Y_j \leq y) \\ &= \lim_{h \to 0+} \binom{n}{j-1, 1, n-j} F(y-h)^{j-1} \Big(\frac{F(y) - F(y-h)}{h}\Big) \big(1 - F(y)\big)^{n-j} \\ &= \binom{n}{j-1, 1, n-j} F(y-h)^{j-1} f(y) \big(1 - F(y)\big)^{n-j}. \end{aligned}$$

Here are some other *equivalent* ways of writing this pdf:

$$f_{Y_j}(y) = \tfrac{n!}{(j-1)!(n-j)!} f(y) F(y-h)^{j-1} \big(1 - F(y)\big)^{n-j}$$

$$f_{Y_j}(y) = \tfrac{\Gamma(n+1)}{\Gamma(j)\Gamma(n+1-j)} f(y) F(y-h)^{j-1} \big(1 - F(y)\big)^{n-j}$$

$$f_{Y_j}(y) = j\tbinom{n}{j} f(y) F(y-h)^{j-1} \big(1 - F(y)\big)^{n-j}$$

$$f_{Y_j}(y) = nf(y)\tbinom{n-1}{j-1} F(y-h)^{j-1} \big(1 - F(y)\big)^{n-j}.$$

Remark.
This pdf agree with the pdfs we found for $j = 1$ and $j = n$ – see page 220 – *Check this*!

Remark.
Although the $X_1, X_2, \ldots, X_n$ may be independent, the Y_j's are definitely *dependent*; in fact, the Y_j will always satisfy

$$Y_1 \leq Y_2 \leq Y_3 \leq \cdots \leq Y_{n-1} \leq Y_n.$$

Notation.
Sometimes the notation $X_{(j)}$ is used to denote the jth ordered statistic instead of Y_j.

Important example.
Suppose $X_1, X_2, \ldots, X_n$ are iid uniform$(0, 1)$. Find the pdf of Y_j.

SOLUTION:
Since $F(y) = y$ and $f(y) = 1$ for $0 < y < 1$, it follows

$$f_{Y_j}(y) = \frac{n!}{(j-1)!(n-j)!} y^{j-1}(1-y)^{n-j} = \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} y^{j-1}(1-y)^{n-j} \quad \text{for } 0 < y < 1,$$

i.e., $Y_j \sim \text{Beta}(j, n+1-j)$.

Application: the sample median
Let $X_1, X_2, \ldots, X_{2n+1}$ be iid r.v.s with PDF $f(x)$ and CDF $F(x)$, and let $Y_1, Y_2, \ldots, Y_{2n+1}$ be the respective ordered statistics. We call Y_{n+1} the ***sample median***. The ***(population) median*** is the value m such that

$$\int_{-\infty}^{m} f(x)\,dx = \frac{1}{2}.$$

Example.
Let $X_1, X_2, X_3 \sim$ iid Exp(1). Then Y_2 is the sample median. Find the PDF of Y_2 and compute $E(Y_2)$.

$$f_{Y_j}(y) = n\,f(y)\binom{n-1}{j-1}[F(y)]^{j-1}[1-F(y)]^{n-j}.$$

$$\begin{aligned} f_{Y_2}(y) &= 3(e^{-y})\binom{2}{1}(1-e^{-y})^{2-1}(e^{-y})^{3-2} \\ &= 6e^{-y}(1-e^{-y})e^{-y} \\ &= 6(e^{-2y}-e^{-3y}) \qquad \text{for } y>0. \end{aligned}$$

$$\begin{aligned} E(y_2) &= \int_0^\infty y\cdot 6(e^{-2y}-e^{-3y})\,dy \\ &= 6\int_0^\infty ye^{-2y}\,dy - 6\int_0^\infty ye^{-3y}\,dy \\ &= 6\left(\frac{1}{2}\right)^2 - 6\left(\frac{1}{3}\right)^2 \\ &= \frac{6}{4}-\frac{6}{9} \\ &= \frac{6\cdot 5}{36} \\ &= \frac{5}{6} \;\approx\; 0.83\overline{3}. \end{aligned}$$

2.2. case of a *general* iid random sample.
In this case $X_1, X_2, \ldots, X_n$ are iid having CDF $F(x)$ not necessarily continuous. We derive the CDF of Y_j by analyzing the event $(Y_j \le y)$. Since the jth smallest is less than or equal to y *is not* saying anything about Y_k for $k > j$, we decompose this event into the union of *mutually exclusive events* as follows:

$$(Y_j \le y) = (Y_n \le y) \cup \bigcup_{k=j}^{n-1} (Y_k \le y, Y_{k+1} > y).$$

Therefore, the CDF of Y_j is

$$\begin{aligned} F_{Y_j}(y) = P(Y_j \le y) &= P(Y_n \le y) + \sum_{k=j}^{n-1} P(Y_k \le y, Y_{k+1} > y) \\ &= F(y)^n + \sum_{k=j}^{n-1} \binom{n}{k} F(y)^k (1-F(y))^{n-k} \\ &= \sum_{k=j}^{n} \binom{n}{k} F(y)^k (1-F(y))^{n-k}. \end{aligned}$$

Remark.
Let's verify that this CDF agrees with the CDFs we found for Y_1 and Y_n on page 220. When $j = n$, the above CDF reduces to

$$\sum_{k=n}^{n} \binom{n}{k} F(y)^k (1-F(y))^{n-k} = F(y)^n.$$

When $j = 1$,

$$\sum_{k=1}^{n} \binom{n}{k} F(y)^k (1-F(y))^{n-k} = 1 - \binom{n}{0} F(y)^0 (1-F(y))^{n-0} = 1 - (1-F(y))^n.$$

Moreover, if the CDF $F(x)$ has pdf $f(x)$, then

$$\begin{aligned} &f_{Y_j}(y) = \tfrac{d}{dy} \textstyle\sum_{k=j}^{n} \binom{n}{k} F(y)^k (1-F(y))^{n-k} \\ &= \textstyle\sum_{k=j}^{n} \binom{n}{k} k F(y)^{k-1} f(y) (1-F(y))^{n-k} - \sum_{k=j}^{n-1} \binom{n}{k} (n-k) F(y)^k f(y) (1-F(y))^{n-k-1} \\ &= \textstyle\sum_{k=j}^{n} \binom{n}{k} k F(y)^{k-1} f(y) (1-F(y))^{n-k} - \sum_{k=j+1}^{n} \binom{n}{k-1} (n-k+1) F(y)^{k-1} f(y) (1-F(y))^{n-k} \\ &= j\binom{n}{j} F(y)^{j-1} f(y) (1-F(y))^{n-j} \\ &\qquad + \textstyle\sum_{k=j+1}^{n} \underbrace{\left\{ k\binom{n}{k} - (n-k+1)\binom{n}{k-1} \right\}}_{=0} F(y)^{k-1} f(y) (1-F(y))^{n-k} \end{aligned}$$

and, therefore, we obtain the pdf of Y_j in another way:

$$f_{Y_j}(y) = j\binom{n}{j} f(y) F(y)^{j-1} (1-F(y))^{n-j}$$

[cf the pdfs on page 223].

3. The joint distribution of $Y_1, Y_2, \ldots, Y_n$.
The last several pages were all about the univariate marginals (CDFs and, if continuous, pdfs) of iid random samples. In this section we derive the joint pdf of the entire collection of ordered statistics $Y_1, Y_2, \ldots, Y_n$ of an iid random sample $X_1, X_2, \ldots, X_n$. To avoid the issue of random variables taking the same value with a positive probability we will assume in this section that $X_1, X_2, \ldots, X_n$ are jointly continuous.

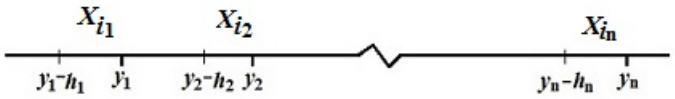
Let $y_1 < y_2 < \cdots < y_n$, $0 < h_i << 1$ for $i = 1, 2, \ldots, n$ and consider the event

$$A := \Big(y_1 - h_1 < Y_1 \leq y_1, y_2 - h_2 < Y_2 \leq y_2, \ldots, y_n - h_n < Y_n \leq y_n\Big).$$

We assume the h_i's are small enough so that the intervals $(y_i - h_i, y_i]$ for $i = 1, 2, \ldots, n$ are pairwise disjoint. Now, the collection of events

$$A_\pi := \Big(y_1 - h_1 < X_{\pi_1} \leq y_1, y_2 - h_2 < X_{\pi_2} \leq y_2, \ldots, y_n - h_n < X_{\pi_n} \leq y_n\Big),$$

for each permutation $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ of the integers $\{1, 2, \ldots, n\}$ are *mutually exclusive*.



Since the X_i's are iid with common CDF $F(x)$ it follows that, for each permutation π,

$$P(A_\pi) = \prod_{i=1}^{n} \Big(F(y_i) - F(y_i - h_i)\Big).$$

Finally, since there are $n!$ such permutations we have

$$\frac{P(y_1 - h_1 < Y_1 \leq y_1, \ldots, y_n - h_n < Y_n \leq y_n)}{h_1 \cdots h_n} = n! \prod_{i=1}^{n} \Big(\frac{F(y_i) - F(y_i - h_i)}{h_i}\Big),$$

and, passing to the limit as each $h_i \to 0+$ we arrive to the

joint pdf of $Y_1, Y_2, \ldots, Y_n$***:***

$$f_{Y_1, Y_2, \ldots, Y_n}(y_1, y_2, \ldots, y_n) = \begin{cases} n! f(y_1) f(y_2) \cdots f(y_n) & \text{for } y_1 < y_2 < \cdots < y_n \\ 0 & \text{elsewhere} \end{cases}.$$

Recap:
When $X_1, X_2, \ldots, X_n \sim$ iid $f(x)$ with CDF$F(x)$, then the ordered statistics of this sample

$$Y_1 \leq Y_2 \leq \cdots \leq Y_n$$

has joint PDF

$$f_{Y_1,Y_2,\ldots,Y_n}\,(y_1, y_2, \ldots, y_n) = \begin{cases} n!\, f(y_1) f(y_2) \ldots f(y_n) & \text{for } y_1 < y_2 < \cdots < y_n \\ 0 & \text{elsewhere} \end{cases}$$

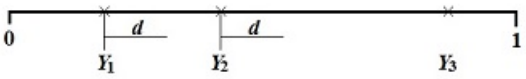
and marginals

$$f_{Y_j}(y) = n\, f(y) \binom{n-1}{j-1} [F(y)]^{j-1} [1 - F(y)]^{n-j}\,.$$

Example (from Sheldon Ross's textbook).
Suppose 3 people are stranded on a 1 mile length of road completely at random. What is the probability that no two people are within a distance d of each other? (Here, $d \leq \frac{1}{2}$ mile).

SOLUTION:
Let X_1, X_2, X_3 represent the location of the three people along the road. We assume $X_1, X_2, X_3 \sim$ iid uniform$(0, 1)$. If Y_1, Y_2, Y_3 are the order statistics of this sample, then no two people are within distance d of each other means that

$$Y_2 - Y_1 > d \quad \text{and} \quad Y_3 - Y_2 > d.$$



$$\begin{aligned}
P(Y_2 - Y_1 > d, Y_3 - Y_2 > d) &= \int_0^{1-2d} \int_{y_1+d}^{1-d} \int_{y_2+d}^{1} 3!\, dy_3\, dy_2\, dy_1 \\
&= 3! \int_0^{1-2d} \int_{y_1+d}^{1-d} (1 - d - y_2)\, dy_2\, dy_1 \qquad \begin{cases} u = 1 - d - y_2 \\ du = -dy_2 \end{cases} \\
&= 3! \int_0^{1-2d} \int_0^{1-2d-y_1} u\, du\, dy_1 \\
&= 3! \int_0^{1-2d} \frac{(1 - 2d - y_1)^2}{2}\, dy_1 \qquad \begin{cases} w = 1 - 2d - y_1 \\ dw = -dy_1 \end{cases} \\
&= \frac{3!}{2} \int_0^{1-2d} w^2\, dw \\
&= (1 - 2d)^3.
\end{aligned}$$

Example.
Let $X_1, X_2, \ldots, X_n \sim$ iid uniform(0,1) r.v.s.
Define the ***spacings***

$$\begin{aligned} W_1 &= Y_1 \\ W_2 &= Y_2 - Y_1 \\ W_3 &= Y_3 - Y_2 \\ &\vdots \\ W_n &= Y_n - Y_{n-1}. \end{aligned}$$

The inverse transformation is

$$\begin{aligned} Y_1 &= W_1 \\ Y_2 &= W_1 + W_2 \\ Y_3 &= W_1 + W_2 + W_3 \\ &\vdots \\ Y_{n-1} &= W_1 + W_2 + W_3 + \cdots + W_{n-1} \\ Y_n &= W_1 + W_2 + W_3 + \cdots + W_{n-1} + W_n. \end{aligned}$$

And so the Jacobian determinant is

$$J = \det \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 1 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & & & & & & \\ 1 & 1 & 1 & 1 & \ldots & 1 & 0 \\ 1 & 1 & 1 & 1 & \ldots & 1 & 1 \end{pmatrix} = 1.$$

If we agree to set $Y_0 = 0$, then, for each integer $i \geq 1$, $Y_{i-1} < Y_i$ implies $0 < W_i < 1$, and $0 < Y_n < 1$ implies $0 < W_1 + W_2 + \cdots + W_n < 1$. The joint pdf of $W_1, W_2, \ldots, W_n$ is

$$f_{W_1,W_2,\ldots,W_n}(w_1, w_2, \ldots, w_n) = n! \quad \text{for } 0 < w_i < 1 \ (i = 1, 2, \ldots, n),\ 0 < w_1 + w_2 + \cdots w_n < 1.$$

4. (bivariate) marginal of Y_i, Y_j where $i < j$
Using an idea similar to how we find the univariate marginals Y_j we can find the bivariate marginal, too. Suppose $X_1, X_2, \ldots, X_n \sim$ iid $f(x)$ (CDF $F(x)$) and consider their ordered statistics $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. Let $1 \leq i < j \leq n$. Let's find $f_{Y_i,Y_j}(y_i, y_j)$ Let $0 < h_i, h_j << 1$.

$P(y_i - h_i < Y_i \leq y_i, \quad y_j - h_j < Y_j \leq y_j) =$
$nf(y_i)h_i \cdot (n-1)f(y_j)h_j \cdot \binom{n-2}{i-1} \cdot [F(y_i - h_i)]^{i-1} \cdot \binom{n-i-1}{n-j} [1-F(y_j)]^{n-j} [F(y_j - h_j) - F(y_i)]^{j-i-1}$

Dividing by $h_1 h_2$ and sending $h_1 \to 0$ and $h_2 \to 0$ gives us

$$f_{Y_i,Y_j}(y_i, y_j) =$$

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!} \cdot [F(y_i)]^{i-1} \cdot [F(y_j) - F(y_i)]^{j-i-1} [1 - F(y_j)]^{n-j} f(y_i) f(y_j).$$

Example.
$X_1, X_2, \ldots, X_n \sim$ iid uniform(0, 1). Find the PDF of the sample range $R = Y_n - Y_1$.

SOLUTION:
Since R is a function of Y_1 and Y_n we need to find the joint pdf of Y_1 and Y_n to start. Using the above computation we find that the joint PDF of Y_1, Y_n is

$$f_{Y_1,Y_n}(y_1, y_n) = n(n-1)(y_n - y_1)^{n-2} \qquad \text{for } 0 < y_1 < y_n < 1.$$

Applying the method of Jabobians we create the variable $S = Y_1$ and find the inverse transformation:

$$y_1 = s \quad \text{and} \quad y_n = r + s$$

having Jacobian: $J = \det \begin{pmatrix} \partial y_1/\partial r & \partial y_1/\partial s \\ \partial y_n/\partial r & \partial y_n/\partial s \end{pmatrix} = \det \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} = -1 \implies |J| = 1.$

The support of the joint density of R and S is

$$0 < r < 1 \text{ and } \ 0 < r + s < 1 \implies 0 < s < 1 - r.$$

Finally,

$$f_{R,S}(r, s) = f_{Y_1,Y_n}(s, r + s)|J| = n(n-1)r^{n-2},$$

and

$$f_R(r) = \int_0^{1-r} n(n-1)r^{n-2}\, dr = n(n-1)r^{n-2}(1-r) \quad \text{for } 0 < r < 1,$$

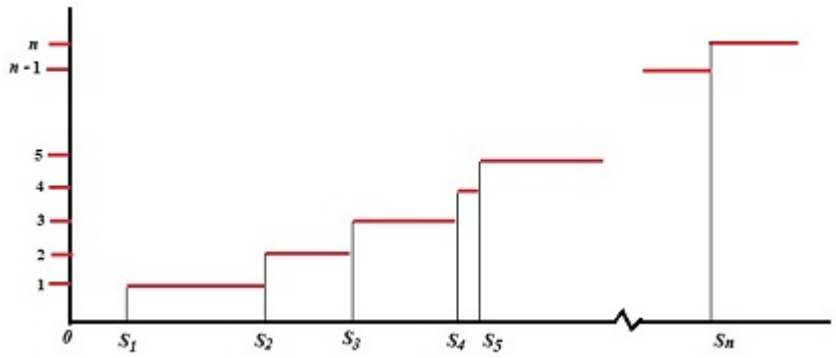
i.e., $R \sim \text{Beta}(n-1, 2)$.

Application (Poisson process).
The stochastic process $(N(t) : t \geq 0)$ such that

1. $P(N(0) = 0) = 1$;
2. For $t > s \geq 0$, $P(N(t) - N(s) = n) = \frac{e^{-\lambda(t-s)}[\lambda(t-s)]^n}{n!}$; i.e.,
$$N(t) - N(s) \sim \text{Poisson}(\lambda(t-s));$$
3. For any $0 \leq t_1 \leq t_2 \leq t_3 \leq \ldots$, all the elements in the following collection
$$N(t_1), N(t_2) - N(t_1), N(t_3) - N(t_2), \ldots$$
are independent;

is called a (time-homogeneous) Poisson process with rate λ.

We let S_n be the time at which the process first reaches the value n: $N(S_n) = n$ and $N(t) < n$ for $t < S_n$. S_n is called the ***arrival time*** of the nth Poisson event.



Let's show that each $S_n \sim \text{Gamma}(n, \frac{1}{\lambda})$.

We use the following important relationship between $N(t)$ and S_n:

$$(N(t) \leq n-1) = (S_n > t).$$

This is true because $S_n > t$ means that the nth Poisson arrival did not occur by time t, i.e., $N(t) \leq n-1$. Therefore, the CDF of S_n is

$$\begin{aligned}
F_{S_n}(t) &= 1 - P(S_n \leq t) \\
&= 1 - P(N(t) \leq n-1) \\
&= 1 - \sum_{j=0}^{n-1} P(N(t) = j) \\
&= 1 - \sum_{j=0}^{n-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!} \\
&= 1 - e^{-\lambda t}\left(1 + \lambda t + \frac{\lambda^2 t^2}{2!} + \cdots + \frac{\lambda^{n-1} t^{n-1}}{(n-1)!}\right).
\end{aligned}$$

Taking a derivative in t, we have the pdf of S_n:

$$f_{S_n}(t) = \lambda e^{-\lambda t}\left(1 + \lambda t + \tfrac{\lambda^2 t^2}{2!} + \cdots + \tfrac{\lambda^{n-1}t^{n-1}}{(n-1)!}\right) - e^{-\lambda t}\left(\lambda + \lambda^2 t + \tfrac{\lambda^3 t^2}{2!} + \cdots + \tfrac{\lambda^{n-1}t^{n-2}}{(n-2)!}\right)$$

and we arrive at

$$f_{S_n}(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!},$$

which is the pdf of a Gamma$(n, \frac{1}{\lambda})$.

Remark.
An interesting fact about the arrival times of a Poisson process is that conditioned on knowing the number of arrivals by time T, the Poisson arrivals occur at times that are uniformly distributed on the interval $[0, T]$, i.e., $(S_1, S_2, \ldots, S_n) \mid N(T) = n$ has the same distribution as the ordered statistics of a uniform$(0, T)$ random sample: $(Y_1, Y_2, \ldots, Y_n)$. Let's show this when $n = 1$, i.e. when $N(T) = 1$. We will show

$$P(S_1 \le t \mid N(T) = 1) = \frac{t}{T},$$

the CDF of a uniform$(0, T)$.

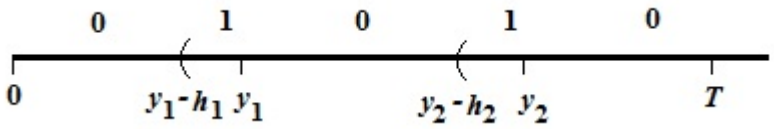
$$\begin{aligned} P(S_1 \le t \mid N(T) = 1) &= \frac{P(S_1 \le t, N(T) = 1)}{P(N(T) = 1)} \\ &= \frac{P(N(T) = 1, N(T) - N(t) = 0)}{P(N(T) = 1)} \\ &= \frac{P(N(T) = 1)P(N(T) - N(t) = 0)}{P(N(T) = 1)} \\ &= \frac{e^{\lambda t}\lambda t \cdot e^{-\lambda(T-t)}}{e^{\lambda T}\lambda T} = \frac{t}{T}. \end{aligned}$$

How about $(S_1, S_2) \mid N(T) = 2$?

We will instead compute this conditional pdf in the same way we approached finding the bivariate marginal pdf of two ordered statistics, i.e., via the limit of

$$\frac{P(t_1 - h_1 < S_1 \le t_1, t_2 - h_2 < S_2 \le t_2 \mid N(T) = 2)}{h_1 h_2}$$

as h_1, h_2 tend to 0.



$$
\begin{aligned}
&P(t_1 - h_1 < S_1 \le t_1, t_2 - h_2 < S_2 \le t_2 \mid N(T) = 2) \\
&= \frac{P(t_1 - h_1 < S_1 \le t_1, t_2 - h_2 < S_2 \le t_2, N(T) = 2)}{P(N(T)) = 2)} \\
&= \frac{P(A \cap B \cap C \cap D \cap E)}{P(N(T) = 2)},
\end{aligned}
$$

where

$$
\begin{aligned}
A:&\ N(t_1 - h_1) = 0, \\
B:&\ N(t_1) - N(t_1 - h_1) = 1, \\
C:&\ N(t_2 - h_2) - N(t_1) = 0, \\
D:&\ N(t_2) - N(t_2 - h_2) = 1, \\
E:&\ N(T) - N(t_2) = 0
\end{aligned}
$$

are all independent by the independent increment property. Now since

$$
\begin{aligned}
&P(A) = P(N(t_1 - h_1) = 0) = e^{-\lambda(t_1 - h_1)}, \\
&P(B) = P(N(t_1) - N(t_1 - h_1) = 1) = e^{-\lambda h_1}\lambda h_1, \\
&P(C) = P(N(t_2 - h_2) - N(t_1) = 0) = e^{-\lambda(t_2 - t_1 + h_2)}, \\
&P(D) = P(N(t_2) - N(t_2 - h_2) = 1) = e^{-\lambda h_2}\lambda h_2, \\
&P(E) = P(N(T) - N(t_2) = 0) = e^{-\lambda(T - t_2)},
\end{aligned}
$$

we have, for $0 < t_1 < t_2 < T$,

$$
\begin{aligned}
&\frac{P(t_1 - h_1 < S_1 \le t_1, t_2 - h_2 < S_2 \le t_2 \mid N(T) = 2)}{h_1 h_2} \\
&\quad = \frac{e^{-\lambda(t_1 - h_1)} e^{-\lambda h_1} \lambda h_1 e^{-\lambda(t_2 - t_1 + h_2)} e^{-\lambda h_2} \lambda h_2 e^{-\lambda(T - t_2)}}{h_1 h_2 e^{-\lambda T}\left(\frac{\lambda^2 T^2}{2!}\right)} = \frac{2!}{T^2}.
\end{aligned}
$$

and, therefore, as h_1 and h_2 tend to 0, the conditional pdf of S_1, S_2 given $N(T) = 2$ is

$$
f_{S_1, S_2 | N(T)}(t_1, t_2 | 2) = 2! \cdot \frac{1}{T} \cdot \frac{1}{T},
$$

which is the joint pdf of the ordered statistics Y_1, Y_2 of a uniform$(0, T)$ random sample.

This last argument can be generalized in a straightforward way to show that

$$
(S_1, S_2, \ldots, S_n) \mid N(T) = n
$$

has the joint pdf $\frac{n!}{T^n}$ for $0 < t_1 < t_2, \cdots < t_n < T$ which is the same distribution as the ordered statistics $Y_1, Y_2, \ldots, Y_n$ of a uniform$(0, T)$ random sample.

Symmetry and Exchangeability

Exchangeable random variables.
$X_1, X_2, \ldots, X_n$ are ***exchangeable*** if for every permutation $i_1, i_2, \ldots, i_n$ of $1, 2, \ldots, n$, we have that $(X_{i_1}, X_{i_2}, \ldots, X_{i_n})$ has the same distribution as $(X_1, X_2, \ldots, X_n)$. For example, when $n = 3$, X_1, X_2, and X_3 are exchangeable if

$$\begin{array}{lll} (X_1, X_2, X_3), & (X_1, X_3, X_2), & (X_2, X_1, X_3), \\ (X_2, X_3, X_1), & (X_3, X_1, X_2), & (X_3, X_2, X_1) \end{array}$$

all have ***the same*** joint distribution.

Symmetric functions.
A real-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is called ***symmetric*** if for every permutation $i_1, i_2, \ldots, i_n$ of $1, 2, \ldots, n$ we have

$$f(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) = f(x_1, x_2, \ldots, x_n) \qquad \text{for all } (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n.$$

Here are some symmetric functions:

1. $f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i, \quad$ where $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$
2. $f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} x_i, \quad$ where $x_1 > 0, x_2 > 0, \ldots, x_n > 0.$

Example.
Which (if any) are symmetric functions?

1. $f(x, y) = \begin{cases} e^{-x-y} & \text{for } x > 0, y > 0 \\ 0 & \text{elsewhere} \end{cases}$

2. $f(x, y, z) = \begin{cases} xy + z & \text{for } x, y, z > 0 \\ 0 & \text{elsewhere} \end{cases}$

3. $f(x, y) = \begin{cases} 2xy & \text{for } 0 < x < 1,\ 0 < y < 2 \\ 0 & \text{elsewhere} \end{cases}$

SOLUTION:

1. This function is symmetric since $-x - y = -y - x$ for all $x > 0, y > 0$.
2. This function is not symmetric. For example, $f(0, 1, 2) = 2 \neq 1 = f(0, 2, 1)$.
3. This function is not symmetric, since the domain for x and y are not identical. For example, $f(0.9, 1.9) = 3.42 \neq 0 = f(1.9, 0.9)$.

Remark.
As the last example demonstrates, in order for a function to be symmetric *both* its rule and its support need to be permutation invariant. The second function has a support that is permutation invariant but its rule is not; whereas, the third function has a rule that is permutation invariant but its support is not.

The importance of symmetric functions follows from this...

Alternate definition of exchangeability.
If $X_1, X_2, \ldots, X_n$ are jointly discrete random variables with joint pmf p, then they are exchangeable if and only if p is a symmetric function.
If $X_1, X_2, \ldots, X_n$ are jointly continuous random variables with joint pdf f, then they are exchangeable if and only if f is a symmetric function.

Some consequences of exchangeability.
If $X_1, X_2, \ldots, X_n$ are exchangeable, then for any $1 \leq k \leq n$, $(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$ has the same distribution as $(X_1, X_2, \ldots, X_k)$.

When $k = 1$, this says that $X_1, X_2, X_3, \ldots, X_n$ all have the same distribution: *exchangeable r.v.s are identically distributed.*

When $k = 2$, this says, for instance,

$$(X_1, X_2), (X_2, X_1), (X_1, X_3), (X_3, X_1), (X_2, X_3), (X_5, X_7), \ldots$$

all have the same distribution.
For example, if $X_1, X_2, \ldots, X_n$ are exchangeable, then

$$E[g(X_3, X_8)] = E[g(X_1, X_2)] = E[g(X_2, X_1)] = \ldots$$

provided that the expectation exists. In fact,

$$E[h(X_1, X_2, \ldots, X_k)] = E[h(X_{i_1}, X_{i_2}, \ldots, X_{i_k})]$$

for any k-tuple $(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$ of $(X_1, X_2, \ldots, X_n)$. ***The expectation will not change under permutations of its arguments.***

If $X_1, X_2, \ldots, X_n$ are exchangeable, then every subcollection of these is also exchangeable.

An important example of an exchangeable sequence is sampling without replacement...

Theorem.
Suppose we have a finite population of N distinct objects comprised of t types. Assume there are n_k objects of type k in this population for $k = 1, 2, \ldots, t$, where $\sum n_k = N$. The experiment is to draw all N objects without replacement. On the ith draw we observe the rv X_i, where $X_i = k$ if we draw a type k object at the ith draw. Then, the sequence $X_1, X_2, \ldots, X_N$ is exchangeable.

Proof.
The sample space Ω of this experiment is the set of all permutations of N objects taken N at a time (assumed equally-likely) $|\Omega| = N!$, and, if $\omega \in \Omega$, then

$$X_i(\omega) = k \text{ iff there is a type } k \text{ object in the } i \text{ entry of } \omega.$$

Consequently,

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N) = \frac{n_1! n_2! \cdots n_t!}{N!}$$

for every sequence $x_1, x_2, \ldots, x_N$ containing exactly n_k $k's$ for $k = 1, 2, \ldots, t$, and it equals 0 otherwise. Notice the joint pmf of $X_1, X_2, \ldots, X_N$ is a symmetric function of $x_1, x_2, \ldots, x_N$, and thus, the collection is *exchangeable*. □

Example.
Suppose we have 4 green balls and 6 blue balls and we draw without replacement. Let $X_i = 1$ if the ith draw is green, $X_i = 0$ if otherwise. Then, from the theorem, $X_1, X_2, \ldots, X_{10}$ are exchangeable. Compute
(a) $P(X_6 = 0 | X_{10} = 1)$
(b) $E[X_6 + (X_9 + X_{10})^3]$

SOLUTION:
(a)
$P(X_6 = 0 | X_{10} = 1) = \frac{P(X_6=0, X_{10}=1)}{P(X_{10}=1)} = \frac{P(X_2=0, X_1=1)}{P(X_1=1)} = P(X_2 = 0 | X_1 = 1) = \frac{6}{9} = \frac{2}{3}$.
Also,
$P(X_6 = 0 | X_{10} = 1) = \frac{P(X_6=0, X_{10}=1)}{P(X_{10}=1)} = \frac{P(X_2=1, X_1=0)}{P(X_1=1)} = \frac{\frac{6}{10} \cdot \frac{4}{9}}{\frac{4}{10}} = \frac{2}{3}$.
(b)

$$\begin{aligned} E[X_6 + (X_9 + X_{10})^3] &= E(X_6) + E[(X_9 + X_{10})^3] = E(X_1) + E[(\underbrace{X_1 + X_2}_{\text{a hypergeom.}})^3] \\ &= \frac{4}{10} + \left(0^3 \cdot \frac{\binom{4}{0}\binom{6}{2}}{\binom{10}{2}} + 1^3 \cdot \frac{\binom{4}{1}\binom{6}{1}}{\binom{10}{2}} + 2^3 \cdot \frac{\binom{4}{2}\binom{6}{0}}{\binom{10}{2}} \right) \\ &= 2. \end{aligned}$$

Example.
Each card in a well-shuffled deck is turned over one after the other.

(a) What's the probability that the last card is a king?

(b) What's the probability that the last two cards are kings?

(c) If the last two cards are kings, what's the probability that the first card is a king?

(d) What's the probability that all the kings happen before all the aces?

solution.
(a) Let a king be a type 1. $P(X_{52} = 1) = P(X_1 = 1) = \frac{1}{13}$.

(b) $P(X_{51} = 1, X_{52} = 1) = P(X_1 = 1, X_2 = 1) = P(X_1 = 1)P(X_2 = 1|X_1 = 1) = \frac{1}{13} \cdot \frac{3}{51}$.

(c) $P(X_1 = 1|X_{51} = 1, X_{52} = 1) = \frac{P(X_1=1,X_{51}=1,X_{52}=1)}{P(X_{51}=1,X_{52}=1)} = \frac{P(X_1=1,X_2=1,X_3=1)}{P(X_1=1,X_2=1)} = \frac{\frac{1}{4}\cdot\frac{3}{51}\cdot\frac{2}{50}}{\frac{1}{4}\cdot\frac{3}{51}} = \frac{1}{25}$.

(d) Let an ace be a type 2. We first compute

$$P(X_{i_1} = 1, X_{i_2} = 1, X_{i_3} = 1, X_{i_4} = 1, X_{i_5} = 2, X_{i_6} = 2, X_{i_7} = 2, X_{i_8} = 2) \tag{8}$$

when $1 \le i_1 < i_2 < i_3 < i_4 < i_5 < i_6 < i_7 < i_8 \le 52$. By exchangeability, for each such ordering of the indices, (8) is equal to

$$P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 2, X_6 = 2, X_7 = 2, X_8 = 2) = \frac{4!4!}{(52)_8}. \tag{9}$$

But, since there are $\binom{52}{8}$ ways to choose the indices in this fashion, the answer is

$$\binom{52}{8} \cdot \frac{4!4!}{(52)_8} = \frac{4!4!}{8!}.$$

The intuition here is that only the relative positions of the 8 cards (the 4 kings and 4 aces) really matter: there are 4!4! orderings that keep the 4 kings before the 4 aces out of 8! equally likely possibilities.

Remark.
A special case of the Theorem is when the population consists of $t = 2$ types, say, successes and failures, and we set $X_i = 1$ if the ith draw is a success and $X_i = 0$ if the ith draw is a failure. Then, for every fixed n, $X_1, X_2, \ldots, X_n$ is exchangeable. This is the hypergeometric experiment of page 100. Another special case of the Theorem is when the population consists of $t = N$ types, i.e., one each of N different types. Again, for every fixed n, $X_1, X_2, \ldots, X_n$ is exchangeable. When $N = 5$ and $n = 3$, this is the example on page 19.

Another important example of exchangeable random variables are iid random variables.

Theorem. (iid random variables are exchangeable)
If $X_1, X_2, \ldots, X_n \sim$ iid, then $X_1, X_2, \ldots, X_n$ are exchangeable.

Remark.
It should be immediately clear why: for instance, if $X_1, X_2, \ldots, X_n$ are iid continuous, then the joint pdf of $X_1, X_2, \ldots, X_n$ is

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n),$$

which is a symmetric function.

Remark.
Suppose $X_1, X_2, \ldots, X_n$ are exchangeable. If we have an event involving some (or all) of these r.v.s, AND, if all the events with the random variables permuted form an exhaustive collection, then often this leads to simple computations of probabilities.

Example. Suppose $X_1, X_2, X_3 \sim$ iid uniform(0, 1). Compute $P(X_3 < X_2)$.

SOLUTION:
Let a denote $P(X_3 < X_2)$. For any continuous distribution, we have

$$a = P(X_3 < X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{x_2} f(x_2)f(x_3)\, dx_3\, dx_2.$$

However, since X_2, X_3 are exchangeable, $P(X_2 < X_3) = P(X_3 < X_2) = a$. Also,

$$\Omega = (X_2 < X_3) \cup (X_3 < X_2) \cup (X_3 = X_2).$$

Note that since these rvs are continuous, we can "ignore" the event $(X_3 = X_2)$ of probability zero. Therefore we have

$$1 = P(X_2 < X_3) + P(X_3 < X_2) + 0 = 2a,$$

$$P(X_3 < X_2) = a = \frac{1}{2}.$$

(*continued*) Compute $P(X_1 < X_2 < X_3)$.

SOLUTION:
The event $(X_1 < X_2 < X_3)$ by exchangeability has the same probability as the event with the rvs permuted in any rearrangement $(X_{i_1} < X_{i_2} < X_{i_3})$ where (i_1, i_2, i_3) is a permutation of 3 indices from $\{1, 2, 3\}$. Also note that all the events with the rvs permuted form a partition of the entire sample space Ω. Therefore we have

$$3! \cdot P(X_1 < X_2 < X_3) = 1, \quad P(X_1 < X_2 < X_3) = \frac{1}{3!}.$$

Remark.
In the example above, exchangeability will still imply that

$$P(X_3 < 2X_2) = P(X_2 < 2X_3) = P(X_1 < 2X_2) = \dots$$

However, the events $(X_3 < 2X_2)$ and $(X_2 < 2X_3)$ are NOT exhaustive.

Example.
Suppose $X, Y \sim$ iid geom(p). Compute $P(X < Y)$.

SOLUTION:
Let a denote $P(X < Y)$; then $P(Y < X) = P(X < Y) = a$, too. Let b denote $P(X = Y)$. Then since $\Omega = (X < Y) \cup (Y < X) \cup (X = Y)$, we have $1 = a + a + b$. Moreover, unlike the last example, these r.v.s are discrete, so b may not be zero.

$$\begin{aligned}
b &= P(X = Y) = \sum_{x=1}^{\infty} P(X = x, Y = x) = \sum_{x=1}^{\infty} P(X = x)P(Y = x) \\
&= \sum_{x=1}^{\infty} (1-p)^{x-1} \cdot p \cdot (1-p)^{x-1} \cdot p \\
&= \frac{p^2}{(1-p)^2} \sum_{x=1}^{\infty} (1-p)^{2x} \\
&= \frac{p^2}{(1-p)^2} \cdot \frac{(1-p)^2}{1-(1-p)^2} \\
&= \frac{p^2}{1-(1-p)^2} = \frac{p}{2-p}.
\end{aligned}$$

Therefore,

$$P(X < Y) = a = \frac{1-b}{2} = \frac{1-p}{2-p}.$$

Example.
Suppose X_1, X_2, X_3, X_4 has the joint pdf

$$f_{X_1,X_2,X_3,X_4}(x_1, x_3, x_3, x_4) = \begin{cases} \frac{1}{2}(x_1 + x_2 + x_3 + x_4) & \text{for } 0 \le x_i \le 1, i = 1, 2, 3, 4 \\ 0 & \text{otherwise.} \end{cases}$$

$$= \frac{1}{2}(x_1 + x_2 + x_3 + x_4)1_{(0,1)}(x_1)1_{(0,1)}(x_2)1_{(0,1)}(x_3)1_{(0,1)}(x_4).$$

Find the probability that X_1 is the largest, i.e., $P(X_1 > X_2, X_1 > X_3, X_1 > X_4)$.

SOLUTION:
The joint pdf is clearly a symmetric function (permutation invariant rule and support), and therefore, the rvs are exchangeable. It must follow that

$P(X_1 \text{ is the largest}) = P(X_2 \text{ is the largest}) = P(X_3 \text{ is the largest}) = P(X_4 \text{ is the largest}).$

Moreover, the events are exhaustive, one rv will be the largest (ties can be neglected since rvs are continuous), and

$$4 \cdot P(X_1 \text{ is the largest}) = 1, \quad P(X_1 \text{ is the largest}) = \frac{1}{4}.$$

Exchangeable events.
If we have events $A_1, A_2, \ldots, A_n$ such that the indicators $X_i = 1_{A_i}$ are exchangeable, then we call the events exchangeable. For example, in sampling without replacement from a finite population of successes and failures, the event $A_i (i = 1, \ldots, n)$ of drawing a success on ith draw are exchangeable events.

Next we introduce a big theorem in exchangeable rvs (students are not responsible for this material in our course).

de Finetti's theorem.
Let $A_1, A_2, \ldots, A_n$ be exchangeable events and let S count the number of these events that occur. Then there exists a unique non-negative function f on $[0, 1]$ such that

$$\int_0^1 f(\theta)\, d\theta = 1,$$

and, for any k such that $0 \le k \le n$,

$$P(S = k) = \int_0^1 \binom{n}{k} \theta^k (1 - \theta)^{n-k} f(\theta)\, d\theta$$

Application: Pólya's urn.
Fix an integer $c \geq -1$. Suppose an urn initially contains a white balls and b black balls ($a+b$ balls total). A ball is drawn uniformly at random from these and replaced together with c more balls of the same color (so that the urn will have $a+b+c$ balls in total at the next draw), and this process is repeated.

Let A_i be the event a white ball is drawn on the ith trial. Then set $X_i = 1_{A_i}$, and $S_n = \sum_{i=1}^{n} X_i$ is the total number of white balls drawn in first n trials.

William Feller in his famous book *Probability Theory and its Applications* (Volume 1 (1968)) showed:

For every n, the events $A_1, A_2, A_3, \ldots, A_n$ are exchangeable, and

$$P(S_n = k) = \binom{n}{k} \frac{a(a+c)\ldots(a+(k-1)c)\cdot b(b+c)\ldots(b+(n-k-1)c)}{(a+b)(a+b+c)\ldots(a+b+(n-1)c)}.$$

This is called the ***Pólya-Eggenberger distribution*** with parameters a, b, and c.

Some special cases of Pólya's urn:
When $c = 0$, this is a binomial experiment and, for every n, $S_n \sim \text{binom}\left(n, \frac{a}{a+b}\right)$.
When $c = -1$, this is a hypergeometric experiment, accordingly, S_n will have a hypergeometric distribution with $N = a+b$, $M = a$.
When $a = b = c = 1$, the distribution reduces to

$$P(S_n = k) = \binom{n}{k} \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}.$$

This is the discrete uniform distribution on $\{0, 1, 2, \ldots, n\}$.
By de Finetti's theorem, we know there exists a pdf $f(\theta)$ such that

$$\frac{1}{n+1} = \binom{n}{k} \int_0^1 \theta^k (1-\theta)^{n-k} f(\theta)\, d\theta \qquad \text{for each } k, \quad 0 \leq k \leq n. \tag{10}$$

So say we take $k = n$, then (10) reduces to

$$\frac{1}{n+1} = \int_0^1 \theta^n f(\theta)\, d\theta.$$

We get lucky here because by inspection we see $f(\theta) = 1$ for $0 < \theta < 1$. And, therefore, for our Polya urn with $a = b = c = 1$, it follows from de Finetti's theorem, for each $k = 0, 1, \ldots, n$,

$$P(S_n = k) = \binom{n}{k} \int_0^1 \theta^k (1-\theta)^{n-k}\, d\theta.$$

The student may want to compare what we did here with the example on page 212.

Remark.
For general a, b, and c the required $f = f(\theta)$ from de Finetti's theorem turns out to be

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \text{for } 0 < \theta < 1.$$

where

$$\alpha = \frac{a}{c} \quad \text{and} \quad \beta = \frac{b}{c}.$$

This says the "mixing" distribution is beta$\left(\frac{a}{c}, \frac{b}{c}\right)$, and

$$\begin{aligned} P(S_n = k) &= \binom{n}{k}\int_0^1 \theta^k(1-\theta)^{n-k} f(\theta)\, d\theta \\ &= \binom{n}{k}\frac{\Gamma(\frac{a+b}{c})}{\Gamma(\frac{a}{c})\Gamma(\frac{b}{c})}\int_0^1 \theta^{k+\frac{a}{c}-1}(1-\theta)^{n-k+\frac{b}{c}-1}\, d\theta \\ &= \binom{n}{k}\frac{\Gamma(\frac{a+b}{c})}{\Gamma(\frac{a}{c})\Gamma(\frac{b}{c})}\frac{\Gamma(k+\frac{a}{c})\Gamma(n-k+\frac{b}{c})}{\Gamma(n+\frac{a+b}{c})}. \end{aligned}$$

VI. First- and second-order results.

Expectations: linearity of expectation.

Assume $X_1, X_2, \ldots, X_n$ are r.v.s possessing expected values. Then

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i).$$

Proof.
We'll assume the rvs are jointly continuous (essentially the same proof works when the rvs are jointly discrete). Proceeding by induction we suppose $n = 2$. Then

$$\begin{aligned}
E(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f_{X,Y}(x,y)\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y \cdot f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx + \int_{-\infty}^{\infty} y \cdot f_Y(y)\,dy = E(X) + E(Y).
\end{aligned}$$

Suppose, for some $n = k \geq 2$, $E\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} E[X_i]$. Then,

$$\begin{aligned}
E\left[\sum_{i=1}^{k+1} X_i\right] &= E\left[\left(\sum_{i=1}^{k} X_i\right) + X_{k+1}\right] \\
&= E\left[\sum_{i=1}^{k} X_i\right] + E(X_{k+1}) && \text{by our base case } k = 2; \\
&= \sum_{i=1}^{k} E(X_i) + E(X_{k+1}) && \text{by our inductive hypothesis;} \\
&= \sum_{i=1}^{k+1} E(X_i),
\end{aligned}$$

which completes the induction. □

Remark.
Whenever we can recognize a random variable as a sum of random variables, then linearity of expectation is often a useful approach to computing its expected value.

Example.
If $X \sim \text{binomial}(n,p)$, then we know $X = \sum_{i=1}^{n} X_i$, where $X_1, X_2, \ldots, X_n \sim \text{iid Bernoulli}(p)$. Since $E(X_i) = p$ for all i, it follows that

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} p = np.$$

Example.
If $X \sim$ hypergeometric (the population has M successes, $N - M$ failures, and we draw a random sample of size n *without* replacement), then

$$X = \sum_{i=1}^{n} X_i \quad \text{where } X_1, X_2, \ldots, X_n \sim \text{Bernoulli}\left(\frac{M}{N}\right).$$

Note that these Bernoullis are dependent and exchangeable. In particular, $E(X_i) = E(X_1) = \frac{M}{N}$ for every i. Therefore we have

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \frac{M}{N} = n \cdot \frac{M}{N}.$$

Remark.
We can also compute $E(X) = \sum_{k=0}^{n} k \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$, but this is brutal in comparison to the previous method.

Linearity of expectation can be useful when trying to compute the expected value of a "counting random variable" since it may be the case that we can decompose the rv into a sum of "simpler" rvs. We illustrate with the next few examples.

Example.
We deal 5 cards from a deck of 52. Compute the expected number of suits in your hand.

SOLUTION: Let X count the number of suits in a hand. Additionally, for $i \in \{1, 2, 3, 4\}$, let

$$X_i = \begin{cases} 1 & \text{if we select (at least) one suit } i \\ 0 & \text{if not.} \end{cases}$$

Then $X = X_1 + X_2 + X_3 + X_4$, and $E(X) = E(X_1) + E(X_2) + E(X_3) + E(X_4)$.

$$E(X_i) = P(X_i = 1) = 1 - P(X_i = 0) = 1 - \frac{\binom{39}{5}}{\binom{52}{5}} \qquad \text{for } i = 1, 2, 3, 4.$$

So

$$E(X) = 4 \cdot \left(1 - \frac{\binom{39}{5}}{\binom{52}{5}}\right) \approx 3.11\ldots$$

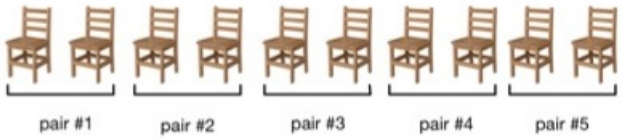
Example.
10 children comprised of 4 boys and 6 girls are randomly seated in 5 pairs of chairs. E.g, if the children are

$$b_1, b_2, b_3, b_4, \quad g_1, g_2, g_3, g_4, g_5, g_6,$$

then

$$(g_3, g_2, b_2, g_1, g_6, b_4, b_1, g_5, g_4, b_3)$$

is one such seating. There are 10! seatings in total.



Let X be the random variable that counts the number of pairs that are occupied by the same gender. Compute $E(X)$.

SOLUTION:
For $i \in \{1, 2, 3, 4, 5\}$, let

$$X_i = \begin{cases} 1 & \text{if pair } i \text{ has 2 children of the same gender} \\ 0 & \text{if otherwise.} \end{cases}$$

Then $X_i \stackrel{\mathscr{D}}{=} X_1$ for all i. Therefore, $X = \sum_{i=1}^5 X_i$ and $E(X) = \sum_{i=1}^5 E(X_i) = 5 \cdot E(X_1)$. We can compute

$$E(X_1) = P(X_1 = 1) = \frac{\binom{4}{2} + \binom{6}{2}}{\binom{10}{2}} = \frac{6 + 15}{45} = \frac{7}{15},$$

so

$$E(X) = \sum_{i=1}^5 E(X_i) = 5 \cdot E(X_1) = \frac{7}{3}.$$

Remark.
With the last two examples we were able to decompose the rv of interest into a *finite* sum of (exchangeable) Bernoullis. This worked especially well because the counting rv was bounded. The next example illustrates a situation where the counting rv of interest is unbounded.

Example.
We have a fair 6-sided die. Compute the expected number of rolls to see every face of the die (at least once).

SOLUTION:

- Let X_1 be the number of rolls to see the first face ($P(X_1 = 1) = 1$).
- Let X_2 be the number of additional rolls to see a face distinct from the first face:

$$X_2 \sim \text{geometric}(\frac{5}{6}), \quad E(X_2) = \frac{1}{\frac{5}{6}} = \frac{6}{5}.$$

- Let X_3 be the number of additional rolls to see a face distinct from the first 2 faces:

$$X_3 \sim \text{geometric}(\frac{4}{6}), \quad E(X_3) = \frac{6}{4}.$$

- Let X_4 be the number of additional rolls to see a face distinct from the first 3 faces:

$$X_4 \sim \text{geometric}(\frac{3}{6}), \quad E(X_4) = \frac{6}{3}.$$

- Similarly,

$$X_5 \sim \text{geometric}(\frac{2}{6}), \quad E(X_5) = \frac{6}{2},$$
$$X_6 \sim \text{geometric}(\frac{1}{6}), \quad E(X_6) = \frac{6}{1}.$$

Therefore, $X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$, and so

$$\begin{aligned} E(X) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) + E(X_6) \\ &= \frac{6}{6} + \frac{5}{6} + \frac{4}{6} + \frac{3}{6} + \frac{2}{6} + \frac{1}{6} \\ &= \frac{147}{10} = 14.7. \end{aligned}$$

Remark.
We see how linearity of expectation (often combined with exchangeability) leads to efficient computation of expected values. It would be nice if we had a similar result for variances. However,

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) \text{ does not always equal } \sum_{i=1}^{n} \text{Var}(X_i)$$

even when the variances of each X_i exist and are finite. Let's investigate when $n = 2$; i.e. let's compute $\text{Var}(X + Y)$ (tacitly assuming that X, Y possess variances).

$$\begin{aligned} Var(X+Y) &= E\left[(X+Y)^2\right] - \left[E(X+Y)\right]^2 \\ &= E\left[X^2 + 2XY + Y^2\right] - \left[E(X) + E(Y)\right]^2 \\ &= E(X^2) + 2E(XY) + E(Y^2) - \{[E(X)]^2 + 2E(X)E(Y) + [E(Y)]^2\} \\ &= Var(X) + Var(Y) + 2[E(XY) - E(X)E(Y)], \end{aligned}$$

and we see that the variance of a sum differs from the sum of the variances by a term involving the expression $E(XY) - E(X)E(Y)$. In probability this expression is called the covariance between X and Y, which we will now investigate.

Covariance between two random variables.
Let X and Y be jointly distributed random variables that have finite means μ_X, μ_Y and finite variances σ_X^2, σ_Y^2. Then the covariance of X, Y is defined as

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Just as variance has a "computationally friendly" version so does covariance:

$$\begin{aligned} \text{Cov}(X,Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[(X - E(X))(Y - E(Y))] \\ &= E[XY - XE(Y) - E(X)Y + E(X)E(Y)] \\ &= E(XY) - E[XE(Y)] - E[E(X)Y] + E[E(X)E(Y)] \\ &= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y). \end{aligned}$$

$\text{Cov}(X,Y)$ can be positive, negative, or zero. If $\text{Cov}(X,Y) = 0$, we say that X and Y are ***uncorrelated***.

What does covariance measure?
Covariance between X and Y is measuring the amount of linear association between the random variables. Loosely speaking a positive covariance means that if the value of one variable increases (decreases) then, on average, the value of the other also increases (decreases). A negative covariance would mean if the value of one variable increases (decreases) then, on average, the value of the other decreases (increases). A zero covariance means there is *no* linear association. For example, if X is the surface area of a strawberry and Y is the number of seeds on the strawberry, then we might suspect $Cov(X,Y) > 0$. On the other hand, if X is the temperature outside and Y is the number of people taking a leisurely stroll in the park, then we might suspect $Cov(X,Y) < 0$.

Properties of covariance.

1. Covariance is symmetric: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
2. $\text{Cov}(X, X) = E(X^2) - [E(X)]^2 = \text{Var}(X)$.
3. Covariance is a bilinear form:

$$\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, Y\right) = \sum_{i=1}^{n} a_i \text{Cov}(X_i, Y);$$

$$\text{Cov}\left(X, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{j=1}^{m} b_j \text{Cov}(X, Y_j).$$

The bilinear property can also be stated as

$$\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, Y_j).$$

Variance of a sum of random variables

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{n}}_{i \neq j} a_i a_j \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{n}}_{i < j} a_i a_j \text{Cov}(X_i, X_j).
\end{aligned}$$

Proof.
Properties 2 and 3 of covariance combine to give

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) &= \text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} a_j X_j\right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}(X_i, X_j).
\end{aligned}$$

Now consider the following $n \times n$ matrix:

$$\begin{bmatrix}
a_1^2 \text{Cov}(X_1, X_1) & a_1 a_2 \text{Cov}(X_1, X_2) & \dots & a_1 a_n \text{Cov}(X_1, X_n) \\
a_2 a_1 \text{Cov}(X_2, X_1) & a_2^2 \text{Cov}(X_2, X_2) & \dots & a_2 a_n \text{Cov}(X_2, X_n) \\
\vdots & \vdots & \ddots & \vdots \\
a_n a_1 \text{Cov}(X_n, X_1) & a_n a_2 \text{Cov}(X_n, X_2) & \dots & a_n^2 \text{Cov}(X_n, X_n)
\end{bmatrix}.$$

Notice that $\text{Var}(\sum_{i=1}^{n} a_i X_i)$ is the sum of all entries of this matrix.
By property 2 of covariance, each entry on the principal diagonal is $a_i^2 \text{Cov}(X_i, X_i) = a_i^2 \text{Var}(X_i)$. By property 1 of covariance, $a_i a_j \text{Cov}(X_i, X_j) = a_j a_i \text{Cov}(X_j, X_i)$. The result follows directly from these observations. □

Special Cases.

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

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\mathrm{Var}(X_i) + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n}}_{i\neq j}\mathrm{Cov}(X_i,X_j),$$

and when the rvs $X_1, X_2, \ldots, X_n$ are exchangeable,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = n\mathrm{Var}(X_1) + n(n-1)\mathrm{Cov}(X_1,X_2).$$

Remark.
From page 204 when X, Y are independent, $E[h_1(X)h_2(Y)] = E[h_1(X)]E[h_2(Y)]$ assuming expectations exist and are finite, and therefore, if X and Y are independent each possessing a variance, then X and Y are uncorrelated, and $E(XY) - E(X)E(Y) = 0$ and $Var(X+Y) = Var(X) + Var(Y)$. In general, if $X_1, X_2, \ldots, X_n$ are independent (or just pairwise independent) possessing variances then they are uncorrelated and

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i).$$

Remark.
There are uncorrelated random variables that are not independent.

Exercise for the student.
Let X be the discrete rv with pmf $P(X=-1) = P(X=1) = \frac{1}{4}, P(X=0) = \frac{1}{2}$, and set $Y = X^2$.

1. Show X, Y are uncorrelated.
2. Show X, Y are NOT independent.

Example.
Suppose $X \sim \text{binom}(n, p)$. Compute $Var(X)$.

SOLUTION:
We first note that $X = X_1 + X_2 + \cdots + X_n$, where $X_1, X_2, \ldots, X_n \sim$ iid Bernoulli(p). Moreover, $Var(X_i) = p(1-p)$. Thus,

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

Example.
Suppose $X \sim$ hypergeometric (the population has M successes, $N - M$ failures, and we draw a random sample of size n without replacement). Compute $\text{Var}(X)$.

SOLUTION:
We know $X = X_1 + X_2 + \cdots + X_n$, where each $X_i \sim \text{Bernoulli}(\frac{M}{N})$. So,

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n}}_{i \neq j} \text{Cov}(X_i, X_j)$$

and, since the rvs are exchangeable, in fact, We claim that for all $i \neq j$, (since the rvs are identically distributed),

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = n\text{Var}(X_1) + n(n-1)\text{Cov}(X_1, X_2).$$

$$\begin{aligned}
\text{Cov}(X_1, X_2) &= E(X_1X_2) - E(X_1)E(X_2) \\
&= P(X_1 = 1, X_2 = 1) - \frac{M}{N} \cdot \frac{M}{N} \\
&= P(X_2 = 1 | X_1 = 1)P(X_1 = 1) - \left(\frac{M}{N}\right)^2 \\
&= \frac{M-1}{N-1} \cdot \frac{M}{N} - \left(\frac{M}{N}\right)^2 \\
&= \frac{M}{N}\left[\frac{M-1}{N-1} - \frac{M}{N}\right] \\
&= \frac{M}{N}\left(\frac{-N+M}{N(N-1)}\right) = -\frac{M}{N}\left(\frac{N-M}{N(N-1)}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{n} X_i\right) &= n\text{Var}(X_1) + n(n-1)\text{Cov}(X_1, X_2) \\
&= n \cdot \frac{M}{N}\left(\frac{N-M}{N}\right) - n(n-1) \cdot \frac{M}{N}\left(\frac{N-M}{N(N-1)}\right) \\
&= n \cdot \frac{M}{N}\left(\frac{N-M}{N}\right)\left(1 - \frac{n-1}{N-1}\right) \\
&= n \cdot \frac{M}{N} \cdot \frac{N-M}{N} \cdot \frac{N-n}{N-1},
\end{aligned}$$

which is the variance of a binom$(n, \frac{M}{N})$ times a ***finite population correction*** term $\frac{N-n}{N-1}$. When this factor is close to 1 the hypergeometric and binomial are nearly identical distributions!

Example.
n people wearing hats walk into a room. They take off their hats and the hats get mixed. Each person randomly selects a hat. Let X denote the number of people who select their own hat. Compute $E(X)$ and $\text{Var}(X)$.

SOLUTION:
Let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th person selects their own hat} \\ 0 & \text{otherwise.} \end{cases}$$

Then $X = \sum_{i=1}^{n} X_i$, and

$$\begin{aligned}
E(X) &= \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} P(X_i = 1) \\
&= \sum_{i=1}^{n} \frac{1}{n} = n \cdot \frac{1}{n} = 1.
\end{aligned}$$

Also, by the exchangeability of these rvs,

$$\begin{aligned}
\text{Var}(X) &= \sum_{i=1}^{n} \text{Var}(X_i) + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n}}_{i \neq j} \text{Cov}(X_i, X_j) \\
&= n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right) + n(n-1) \cdot \left[\frac{1}{n} \cdot \frac{1}{n-1} - \left(\frac{1}{n}\right)^2\right] \\
&= 1 - \frac{1}{n} + 1 - \frac{n-1}{n} \\
&= 1.
\end{aligned}$$

Note that $E(X) = \text{Var}(X) = 1$.

The Cauchy–Schwarz inequality and correlation.

The Cauchy–Schwarz inequality. If X, Y are random variables having finite means and variances, then

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)} = [E(X^2)E(Y^2)]^{\frac{1}{2}}.$$

Proof. (sketch)
For real c define the function $g(c) = E[(X + cY)^2]$. This function is nonnegative; moreover,

$$g(c) = E(X^2) + 2cE(XY) + c^2E(Y^2).$$

The value of c that minimizes g satisfies $g'(c) = 2E(XY) + 2cE(Y^2) = 0$, namely, $c = -\frac{E(XY)}{E(Y^2)}$. Substituting this value of c into $g(c)$ and noting that $g(c) \geq 0$ gives:

$$E(X^2) - 2\frac{E(XY)^2}{E(Y^2)} + \left(\frac{E(XY)}{E(Y^2)}\right)^2 E(Y^2) \geq 0,$$

which reduces to the Cauchy–Schwarz inequality.

Remark.
If we replace X by the rv $X - \mu_X$ and replace Y by rv $Y - \mu_Y$ in the Cauchy–Schwarz inequality, then we get the equivalent form:

$$|\mathrm{Cov}(X, Y)| \leq [\mathrm{Var}(X)\mathrm{Var}(Y)]^{\frac{1}{2}} = \sigma_X\sigma_Y.$$

Correlation between X and Y.
Define the correlation between X and Y by

$$\rho_{X,Y} = \mathrm{corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X\sigma_Y}.$$

Correlation measures the "strength" of linear association between X and Y. From the previous remark, the Cauchy–Schwarz inequality implies

$$|\rho_{X,Y}| \leq 1.$$

The closer $|\rho_{X,Y}|$ is to 1, the stronger the linear association.

Conditional expectation.

We have several different forms; first for **discrete r.v.s X**.
Conditional expectations where we condition on events:
If A is an event with $P(A) > 0$, then

$$E(X|A) = \sum_x x \cdot P(X = x|A) = \sum_x x \cdot \frac{P(\{X = x\} \cap A)}{P(A)}.$$

Example.
Suppose $X \sim$ geometric$(\frac{1}{3})$, and A is the event a success happens on or before the third trial, i.e. $A = (X \leq 3)$. In this example,

$$P(X \leq 3) = \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} + \frac{1}{3} \cdot \left(\frac{2}{3}\right)^2 = \frac{19}{27}.$$

$$\begin{aligned} E(X|A) &= E(X|X \leq 3) \\ &= \sum_x x \cdot \frac{P(\{X = x\} \cap \{X \leq 3\})}{P(X \leq 3)} \\ &= \frac{1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} \cdot \frac{2}{3} + 3 \cdot \frac{1}{3} \cdot \left(\frac{2}{3}\right)^2}{\frac{19}{27}} \\ &= \frac{33}{19}. \end{aligned}$$

Example.
We toss a fair coin $n = 5$ times. Compute the number of heads given the first and last flips differ in parity (i.e., if the first toss is heads, then the last toss is tails, and conversely). Let A be the event that the first and last flips differ in parity. Then

$$P(A) = \frac{1}{2}.$$

$$\begin{aligned} E(X|A) &= \sum_{x=0}^{5} x \cdot \frac{P(X = x, A)}{P(A)} \\ &= 2 \sum_{x=0}^{5} x \cdot P(X = x, A) \\ &= 2 \cdot (1 \cdot P(X = 1, A) + 2 \cdot P(X = 2, A) + 3 \cdot P(X = 3, A) + 4 \cdot P(X = 4, A)) \\ &= 2 \cdot \left(1 \cdot \frac{2}{32} + 2 \cdot \frac{3+3}{32} + 3 \cdot \frac{3+3}{32} + 4 \cdot \frac{2}{32}\right) \\ &= \frac{80}{32}. \end{aligned}$$

Example. (Cute but specialized)
You flip a fair coin n times.
1. How many sequences have no consecutive heads? Let A_n denote those sequences with no two consecutive heads.
2. Let $X = 1$ if a sequence starts with heads, $X = 0$ otherwise. Compute $E(X)$ and $E(X|A_n)$.

SOLUTION:
1 is tricky. It is best analyzed by recursions. Let's do this now:
If we flip a coin $n = 1$ time, then $|A_1| = 2$ since $A_1 = \{h, t\}$.
If we flip a coin $n = 2$ times, then $|A_2| = 3$ since $A_2 = \{ht, th, tt\}$.
Let $F_n = |A_n|$. We just argued that $F_1 = 2, F_2 = 3$. To analyze general F_n with $n > 2$, we "break up" A_n into two mutually exclusive pieces:

case 1: those sequences of length n that start with t:

$$\overset{t}{\underline{}}\ \underline{}\ \underline{}\ \cdots\ \underline{}$$
$$1\ \ 2\ \ 3\ \ \cdots\ \ n$$

and

case 2: those sequences of length n that start with h:

$$\overset{h}{\underline{}}\ \underline{}\ \underline{}\ \cdots\ \underline{}$$
$$1\ \ 2\ \ 3\ \ \cdots\ \ n$$

In case 1, since the sequences start with "tails", the event of no two consecutive "heads" is the same as the number of sequences of length $n-1$ with no two consecutive heads and there are F_{n-1} of these. In case 2, if we want no two consecutive heads, then since the sequence starts with heads the next toss will need to be tails which leaves $n-2$ tosses for no two consecutive heads and there are F_{n-2} of these. Putting this together we have

$$F_n = F_{n-1} + F_{n-2}$$

for $n = 3, 4, 5, 6, 7, \ldots$, which is the answer to question 1. (F_n) is called a ***Fibonacci sequence***:

n	1	2	3	4	5	6	$\cdots$
F_n	2	3	5	8	13	21	$\cdots$

Moreover, $E(X) = \frac{1}{2}$ and

$$E(X|A_n) \ = \ 0 \cdot P(X = 0|A_n) + 1 \cdot P(X = 1|A_n) = \frac{F_{n-2}}{F_n}.$$

FYI: Here are the 13 sequences when $n = 5$.
those that start with t:

$$ththt, thtth, thttt, tthth, tthtt, tttht, ttttt, ttttt$$

and those that start with h:

$$hthth, hthtt, httht, htttt, htttt.$$

Conditional expectation with a conditioning rv.
A common situation of conditional expectation is when the conditioning event involves another random variable as follows:

$$\begin{aligned} E(X|Y=y) &= \sum_x x \cdot P(X=x|Y=y) \quad \text{when } X,Y \text{ jointly discrete;} \\ &= \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\,dx \quad \text{when } X,Y \text{ jointly continuous.} \end{aligned}$$

Notice in either case that $E(X|Y=y)$ is a function of y : $E(X|Y=y) = g(y)$.
However, if X and Y are independent, it is clear that for all y, $E(X|Y=y) = E(X)$.

More generally, the idea here is that we are taking the expected value of X with respect to the ***conditional distribution*** of X given $Y=y$.

Example.
Suppose X, Y has joint pmf (marginals shown for convenience):

	$y=0$	$y=1$	$y=2$	$y=3$	$p_X(x)$
$x=1$	.1	.1	.1	0	.3
$x=2$	0	.2	.1	.1	.4
$x=3$	.1	.1	0	.1	.3
$p_Y(y)$	.2	.4	.2	.2	

Compute $E(X|Y=y)$ for each y and $E(Y|X=x)$ for each x.

SOLUTION:
I'll construct the conditional pmfs within the given table to illustrate.
From our table, to compute $p_{X|Y}(x|y) = P(X=x|Y=y) = \frac{p(x,y)}{p_Y(y)}$ we just take each entry and normalize by the *column* sum (i.e., the marginal of Y at y):

	$y=0$	$y=1$	$y=2$	$y=3$
$x=1$	$.1\ \frac{.1}{.2} = \frac{1}{2}$	$.1\ \frac{.1}{.4} = \frac{1}{4}$	$.1\ \frac{.1}{.2} = \frac{1}{2}$	$0\ \frac{0}{.2} = 0$
$x=2$	$0\ \frac{0}{.2} = 0$	$.2\ \frac{.2}{.4} = \frac{1}{2}$	$.1\ \frac{.1}{.2} = \frac{1}{2}$	$.1\ \frac{.1}{.2} = \frac{1}{2}$
$x=3$	$.1\ \frac{.1}{.2} = \frac{1}{2}$	$.1\ \frac{.1}{.4} = \frac{1}{4}$	$0\ \frac{0}{.2} = 0$	$.1\ \frac{.1}{.2} = \frac{1}{2}$
$p_Y(y)$	**.2**	**.4**	**.2**	**.2**

$$\begin{aligned} E(X|Y=0) &= 1\cdot\frac{1}{2} + 2\cdot 0 + 3\cdot\frac{1}{2} = 2. \\ E(X|Y=1) &= 1\cdot\frac{1}{4} + 2\cdot\frac{1}{2} + 3\cdot\frac{1}{4} = 2. \\ E(X|Y=2) &= 1\cdot\frac{1}{2} + 2\cdot\frac{1}{2} + 3\cdot 0 = 1.5. \\ E(X|Y=3) &= 1\cdot 0 + 2\cdot\frac{1}{2} + 3\cdot\frac{1}{2} = 2.5. \end{aligned}$$

The conditional pmf of Y given $X = x$ is constructed in a similar manner, we just normalize each entry by the *row* sum (i.e., the marginal of X at x):

	$y=0$	$y=1$	$y=2$	$y=3$	$p_X(x)$
$x=1$	$.1\ \frac{.1}{.3}=\frac{1}{3}$	$.1\ \frac{.1}{.3}=\frac{1}{3}$	$.1\ \frac{.1}{.3}=\frac{1}{3}$	$0\ \frac{0}{.3}=0$	**.3**
$x=2$	$0\ \frac{0}{.4}=0$	$.2\ \frac{.2}{.4}=\frac{1}{2}$	$.1\ \frac{.1}{.4}=\frac{1}{4}$	$.1\ \frac{.1}{.4}=\frac{1}{4}$	**.4**
$x=3$	$.1\ \frac{.1}{.3}=\frac{1}{3}$	$.1\ \frac{.1}{.3}=\frac{1}{3}$	$0\ \frac{0}{.3}=0$	$.1\ \frac{.1}{.3}=\frac{1}{3}$	**.3**

$$\begin{aligned}
E(Y|X=1) &= 0\cdot\frac{1}{3}+1\cdot\frac{1}{3}+2\cdot\frac{1}{3}+3\cdot 0 = 1.\\
E(Y|X=2) &= 0\cdot 0+1\cdot\frac{1}{2}+2\cdot\frac{1}{4}+3\cdot\frac{1}{4}=\frac{7}{4}.\\
E(Y|X=3) &= 0\cdot\frac{1}{3}+1\cdot\frac{1}{3}+2\cdot 0+3\cdot\frac{1}{3}=\frac{4}{3}.
\end{aligned}$$

Example.

Suppose X, Y are jointly continuous with joint pdf $f(x,y) = \begin{cases} e^{-y} & 0<x<y<\infty; \\ 0 & \text{elsewhere.} \end{cases}$.

Compute $E(Y|X=x)$.

SOLUTION:
Since $E(Y|X=x)$ is just the expectation of the conditional distribution of Y given $X=x$, we compute the conditional pdf of Y given $X=x$. When $x>0$ and $y>x$,

$$\begin{aligned}
f_{Y|X}(y|x) &= \frac{f(x,y)}{f_X(x)}\\
&= \frac{e^{-y}}{\int_x^\infty e^{-y}\,dy} = \frac{e^{-y}}{e^{-x}}\\
&= \begin{cases} e^{-(y-x)} & \text{for } y>x; \\ 0 & \text{elsewhere.} \end{cases}
\end{aligned}$$

This is called a ***delayed (unit) exponential*** (the delay, here, is x and, thereafter $Y|X=x$ behaves as an exp(1) distribution).

Now,

$$\begin{aligned}
E(Y|X=x) &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\,dy\\
&= \int_x^\infty y e^{-(y-x)}\,dy \qquad \text{substitution } u=y-x,\ du=dy\\
&= \int_0^\infty (u+x)e^{-u}\,du = \int_0^\infty \underbrace{ue^{-u}}_{\sim \text{Gamma}(2,1)}\,du + x\int_0^\infty e^{-u}\,du\\
&= 1+x.
\end{aligned}$$

Exercise for the student.
Continue with the last example and compute $E(X|Y=y)$ where $y>0$.
Your answer should be $E(X|Y=y)=\frac{y}{2}$.

Conditional expectations can be thought of as random variables.
We learned that the conditional expectation of X given $Y=y$ is just a function of y, namely,

$$E(X|Y=y)=g(y).$$

Therefore, $E(X|Y)$ can be thought of as $g(Y)$, i.e., $E(X|Y)$ is the random variable that returns the value $g(y)$ when $Y=y$.

The following tells us that X and $E(X|Y)$ have the same expected value. This is an especially useful and important result in probability. It allows us to compute the (unconditional) expectation of a random variable by way of a conditional distribution whose expectation is computable.

Property 1: Law of total expectation.

$$E[E(X|Y)]=E(X).$$

Proof. I'll illustrate the proof when X and Y are jointly discrete. We have

$$\begin{aligned}
E[E(X|Y)] &= E[g(Y)] \\
&= \sum_y g(y)\cdot P(Y=y) \\
&= \sum_y \sum_x (x\cdot P(X=x|Y=y))\cdot P(Y=y) \\
&= \sum_x \sum_y x\cdot P(X=x,Y=y) \\
&= \sum_x x \sum_y P(X=x,Y=y) \\
&= \sum_x xP(X=x) \\
&= E(X).
\end{aligned}$$

□

Exercise for the student. Prove this theorem when X and Y are jointly continuous.

Example.
Suppose $X \sim \text{Gamma}(\alpha, \beta)$ and $Y|X \sim \text{Gamma}(X, X)$. Derive $E(Y)$.

SOLUTION:
One approach is to find the unconditional distribution of Y and then use brute force – i.e., $E(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y)dy$. The joint pdf is $f(x,y) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \cdot \frac{y^{x-1}e^{-y/x}}{x^x \Gamma(x)}$ for $x > 0, y > 0$, and, for $y > 0$,

$$f_Y(y) = \int_0^\infty \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \cdot \frac{y^{x-1}e^{-y/x}}{x^x \Gamma(x)}\, dx,$$

which is not easily computable. Therefore, it appears the brute force computation of $E(Y)$ is not the way to go.

However, in this example we can employ the law of total expectation:

$$\begin{aligned} E(Y) &= E[E(Y|X)]. \quad \text{where } Y|X \sim \text{Gamma}(X, X) \\ &= E(X^2) \\ &= \text{Var}(X) + [E(X)]^2 \\ &= \alpha\beta^2 + (\alpha\beta)^2 \\ &= \alpha(\alpha+1)\beta^2. \end{aligned}$$

Example.
Your friend tosses a coin (with probability p of coming up heads) n times. Let X denote the number of heads your friend obtains. Then she hands you the same coin, and you get to toss it X times. Let Y denote the number of heads that you obtain. Compute $E(Y)$.

SOLUTION:
According to the problem, we know $X \sim \text{binom}\,(n, p)$ and $Y|X \sim \text{binom}\,(X, p)$. Therefore, $E(Y|X) = Xp$ and

$$E(Y) = E[E(Y|X)] = E\,(Xp) = pE(X) = p \cdot np = np^2.$$

We close this section by stating (and proving) several other important ***properties that conditional expectations*** enjoy.

Property 2.
Let X and Y be jointly distributed r.v.s. and h be any functions, then,

$$E[Xh(Y)|Y] = h(Y)E(X|Y).$$

This says functions of the given rv can be treated as scalars and, thus, can be factored out.

Proof. (only presented when X and Y are jointly discrete)
Let y be a value of Y.

$$\begin{aligned} E[Xh(Y)|Y=y] &= \sum_x xh(y)P[X=x|Y=y] \\ &= h(y)\sum_x xP[X=x|Y=y] \\ &= h(y)E[X|Y=y] \end{aligned}$$

Therefore,

$$E(Xh(Y)|Y) = h(Y)E(X|Y).$$

□

Property 3
Let U, V, W be any jointly distributed r.v.s. and let a and b be any constants, then

$$E(aU + bV|W) = aE(U|W) + bE(V|W).$$

This says conditional expectation is a linear operation.

Property 4
If X and Y are independent, then $E(X|Y) = E(X)$.

Property 4 should be immediate from the fact that the conditional distribution must equal the unconditional distribution by independence.

Property 5
If c is a constant, then $E(c|Y) = c$.

Conditional variance.

We define the conditional variance as follows:

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

Also,

$$Var(X|Y) = E[(X - E(X|Y)^2|Y],$$

where we think of $Var(X|Y)$ as a random variable.

Remark.
$Var(X|Y)$ is the variance of the conditional distribution of X given Y.

Example.
Suppose X, Y are jointly continuous with joint pdf

$$f(x,y) = xe^{-x(1+y)} \text{ for } x > 0, y > 0.$$

Compute $Var(X|Y)$.

SOLUTION: We first find the conditional density of X given $Y = y$. Fix $y > 0$.

$$f_Y(y) = \int_0^\infty xe^{-x(1+y)}\,dx = (1+y)^{-2}\Gamma(2) = \frac{1}{(1+y)^2},$$

therefore, $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = (1+y)^2 xe^{-x(1+y)}$, i.e., $X|Y = y \sim \text{Gamma}(2, \frac{1}{1+y})$. Consequently, $Var(X|Y) = \frac{2}{(1+y)^2}$.

Law of total variance.

$$Var(X) = E(Var(X|Y)) + Var(E(X|Y)).$$

Proof.

$$\begin{aligned}
Var(X) &= E(\{X - E(X)\}^2) \\
&= E(\{X - E(X|Y) + E(X|Y) - E(X)\}^2) \\
&= \underbrace{E(\{X - E(X|Y)\}^2)}_{(1)} \\
&\quad + \underbrace{E(\{E(X|Y) - E(X)\}^2)}_{(2)} \\
&\quad + \underbrace{E(\{X - E(X|Y)\}\{E(X|Y) - E(X)\})}_{(3)} \\
&= E(Var(X|Y)) + Var(E(X|Y)) + 0,
\end{aligned}$$

where
(1): $E(\{X - E(X|Y)\}^2) = E[E(\{X - E(X|Y)\}|Y] = E[Var(X|Y)]$;
(2): Since $E(X) = E[E(X|Y)]$,

$$E(\{E(X|Y) - E(X)\}^2) = E(\{E(X|Y) - E[E(X|Y)]\}^2 = Var(E(X|Y));$$

(3):

$$\begin{aligned} E(\{X - E(X|Y)\}\{E(X|Y) - E(X)\}) &= E(E(\{X - E(X|Y)\}\{E(X|Y) - E(X)\}|Y) \\ &= E[\{E(X|Y) - E(X)\}E(X - E(X|Y)|Y)] \\ &= 0. \end{aligned}$$

Two exercises.
(1) Show that the constant c that minimizes $E[(X - c)^2]$ is $c = E(X)$
(2) Show that the function $g(Y)$ that minimizes $E[(X - g(Y))^2]$ is $g(Y) = E(X|Y)$.
Hint: start by subtracting and adding $E(X|Y)$ under the square. Use ideas in the proof above.

Example.
A miner is trapped in a room with 3 doors.

- Door 1 leads the miner to freedom in 3 hours.
- Door 2 leads the miner back to the room after 5 hours.
- Door 3 leads the miner back to the room after 7 hours.

The miner is equally likely to choose any of the 3 doors at each trial.
Compute the expected time until the miner is free and variance of the time.

SOLUTION: If $D =$ door chosen, then $P(D = 1) = P(D = 2) = P(D = 3) = \frac{1}{3}$.
Also, if X is the time to freedom, then

$$\begin{aligned} E(X|D = 1) &= 3 \\ E(X|D = 2) &= 5 + E(X) \\ E(X|D = 3) &= 7 + E(X). \end{aligned}$$

Therefore, by the law of total expectation,

$$\begin{aligned} E(X) &= E[E(X|D)] \\ &= E(X|D = 1)P(D = 1) + E(X|D = 2)P(D = 2) + E(X|D = 3)P(D = 3) \\ &= 3 \cdot \frac{1}{3} + (5 + E(X)) \cdot \frac{1}{3} + (7 + E(X)) \cdot \frac{1}{3} \\ &= 1 + \frac{5}{3} + \frac{7}{3} + \frac{2}{3}E(X). \end{aligned}$$

Hence, $\frac{1}{3}E(X) = 5$, meaning $E(X) = 15$ hours.

For $Var(X)$, we first compute $E(X^2)$

$$\begin{aligned} E(X^2) &= E(X^2|D=1)\cdot\frac{1}{3} + E(X^2|D=2)\cdot\frac{1}{3} + E(X^2|D=2)\cdot\frac{1}{3} \\ &= 9\cdot\frac{1}{3} + E[(5+X)^2]\cdot\frac{1}{3} + E[(7+X)^2]\cdot\frac{1}{3} \\ &= 3 + (25+10E(X)+E(X^2))\cdot\frac{1}{3} + (49+14E(X)+E(X^2))\cdot\frac{1}{3} \\ &= 3 + \frac{25}{3} + \frac{49}{3} + (\frac{10}{3}+\frac{14}{3})E(X) + \frac{2}{3}E(X^2). \end{aligned}$$

Therefore, $\frac{1}{3}E(X^2) = \frac{497}{3} \implies E(X^2) = 497$. Thus,

$$Var(X) = 497 - (15)^2 = 272.$$

Example. (***Compound random variables***)
In a given year, the number N of claims is a random variable having a mean μ_N and variance σ_N^2. Let X_i be the size of the ith claim in the year, so that

$$S_N = \sum_{i=1}^{N} X_i = X_1 + X_2 + \cdots + X_N$$

represents the total loss incurred in the year. Moreover, we assume
$X_1, X_2, X_3, \cdots$ are iid with mean μ_X and variance σ_X^2
X_i's are independent of N, and
N has mean μ_N and variance σ_N^2.
Compute the mean and variance of the total loss incurred in the year, i.e., compute $E(S_N)$ and $Var(S_N)$.

SOLUTION:
Notice S_N is a sum of a random number of iid rvs.

$$\begin{aligned} E(S_N) = E[\sum_{i=1}^{N} X_i] &= E\Big[E\big(\sum_{i=1}^{N} X_i|N\big)\Big] \qquad \text{by the law of total expectation} \\ &\overset{(*)}{=} E[N\mu_X] \\ &= \mu_X E(N) \\ &= \mu_X\mu_N \end{aligned}$$

$^{(*)}$ Notice that

$$E\big(\sum_{i=1}^{N} X_i|N=n\big) = E\big(\sum_{i=1}^{n} X_i|N=n\big) = E\big(\sum_{i=1}^{n} X_i\big) = \sum_{i=1}^{n} E(X_i) = n\mu_X = g(n),$$

where the second equality follows by property 4 of conditional expectation. Therefore, $E\big(\sum_{i=1}^{N} X_i|N\big) = g(N) = N\mu_X$.

As for the variance,

$$\begin{aligned} Var(S_N) = Var\Big(\sum_{i=1}^{N} X_i\Big) &= E\Big[\big(\sum_{i=1}^{N} X_i\big)^2\Big] - \mu_X^2\mu_N^2 \\ &\overset{(1)}{=} E\Big(\sum_{i=1}^{N}\sum_{j=1}^{N} X_iX_j\Big) - \mu_X^2\mu_N^2 \\ &\overset{(2)}{=} E(N\sigma_X^2 + N^2\mu_X^2) - \mu_X^2\mu_N^2 \\ &= \mu_N\sigma_X^2 + (\sigma_N^2 + \mu_N^2)\mu_X^2 + \mu_X^2\mu_N^2 \\ &= \mu_N\sigma_X^2 + \sigma_N^2\mu_X^2, \end{aligned}$$

where we used the following facts:

(1):

$$(\sum_{i=1}^{N} X_i)^2 = (\sum_{i=1}^{N} X_i)(\sum_{j=1}^{N} X_j) = \sum_{i=1}^{N}\sum_{j=1}^{N} X_iX_j,$$

and

(2): when $N = n$,

$$\begin{aligned} E[\sum_{i=1}^{N}\sum_{j=1}^{N} X_iX_j|N=n] &= E[(\sum_{i=1}^{n}\sum_{j=1}^{n} X_iX_j|N=n] \\ &= E[\sum_{i=1}^{n}\sum_{j=1}^{n} X_iX_j] \quad \text{(since } X_i\text{'s are independent of } N) \\ &= E[\sum_{i=1}^{n} X_i^2 + \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} X_iX_j] \\ &= \sum_{i=1}^{n} E(X_i^2) + \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} E(X_i)E(X_j) \\ &= n(\sigma_X^2 + \mu_X^2) + n(n-1)\mu_X^2 \\ &= n\sigma_X^2 + n^2\mu_X^2 = g(n), \end{aligned}$$

and, therefore, $E[\sum_{i=1}^{N}\sum_{j=1}^{N} X_iX_j|N] = g(N) = N\sigma_X^2 + N^2\mu_X^2$.

Remark.
We just showed $Var(S_N) = \mu_N\sigma_X^2 + \sigma_N^2\mu_X^2$. The reader should take notice of how large this variance is compared to $E(N)Var(X_1) = \mu_N\sigma_X^2$. The variability in the number of terms in the random sum is contributing a huge amount to the overall variability in the total loss.

The bivariate Normal distribution.
We consider a bivariate distribution which is *jointly* normally distributed. To motivate let's first suppose that $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ are *independent*. If Z_1 and Z_2 are independent standard normal rvs, it follows from the corollary on page 163 that

$$\begin{aligned} X_1 &= \mu_1 + \sigma_1 Z_1 \\ X_2 &= \mu_2 + \sigma_2 Z_2 \end{aligned} \tag{11}$$

and, in fact, the joint pdf is just the product of the marginals of X_1 and X_2:

$$f_{X_1,X_2}(x_1, x_2) = \frac{e^{-\frac{1}{2}\left(\left(\frac{x_1-\mu_1}{\sigma_1^2}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2^2}\right)^2\right)}}{2\pi\sigma_1\sigma_2}.$$

Now, let ρ be a constant, $-1 \le \rho \le 1$, and consider the (linear) transformation

$$\begin{aligned} X_1 &= \mu_1 + \sigma_1 Z_1 \\ X_2 &= \mu_2 + \sigma_2\rho Z_1 + \sigma_2\sqrt{1-\rho^2} Z_2 \end{aligned} \quad = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}. \tag{12}$$

When $\rho = 0$, (12) reduces to (11). Moreover, when ρ is *strictly* between -1 and 1, (12) is one-to-one and the method of Jacobians will show that the joint pdf of X_1 and X_2 is

$$f(x_1, x_2) = \frac{e^{-\frac{1}{2(1-\rho^2)}\cdot\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}},$$

which is the ***bivariate normal pdf*** with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$, and ρ.

Fact:
Let's show that parameters of the bivariate normal have the following meanings:

$$\mu_1 = E(X_1),\ \mu_2 = E(X_2),\ \sigma_1^2 = Var(X_1),\ \sigma_2^2 = Var(X_2),\ \text{and } \rho = \mathrm{corr}(X_1, X_2).$$

From (12) and the corollary on page 163, $\mu_1 = E(X_1)$ and $\sigma_1^2 = Var(X_1)$. Also, by taking the expected value in the second equation it follows that $\mu_2 = E(X_2)$, and by taking variance in the first second equation and noting that Z_1 and Z_2 are independent, $Var(X_2) = \sigma_2^2\rho^2 Var(Z_1) + \sigma_2^2(1-\rho^2)Var(Z_2) = \sigma_2^2$. Finally, using the properties of covariance

$$\begin{aligned} cov(X_1, X_2) &= cov(\mu_1 + \sigma_1 Z_1, \mu_2 + \sigma_2\rho Z_1 + \sigma_2\sqrt{1-\rho^2} Z_2) \\ &= cov(\sigma_1 Z_1, \sigma_2\rho Z_1 + \sigma_2\sqrt{1-\rho^2} Z_2) \\ &= \sigma_1\sigma_2\rho \underbrace{cov(Z_1, Z_1)}_{=Var(Z_1)=1} + \sigma_1\sigma_2\sqrt{1-\rho^2} \underbrace{cov(Z_1, Z_2)}_{=0} \\ &= \sigma_1\sigma_2\rho \end{aligned}$$

from which it follows $\mathrm{corr}(X_1, X_2) = \frac{cov(X_1,X_2)}{\sigma_1\sigma_2} = \rho$. □

Remark.
As mentioned already, if $\rho = 0$, i.e. if X_1 and X_2 are uncorrelated, then the joint pdf becomes

$$f(x_1, x_2) = \frac{e^{-\frac{1}{2(1-\rho^2)}\cdot\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \overset{\rho=0}{=} \frac{e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2}}{\sigma_1\sqrt{2\pi}} \cdot \frac{e^{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2}}{\sigma_2\sqrt{2\pi}},$$

and the joint pdf is the product of its marginal pdf. Thus,

Jointly distributed normals *that are* ***uncorrelated*** *are* ***independent!***

Remark.
When $|\rho| = 1$, then the transformation (12) is *not* one-to-one. In fact, when, say, $\rho = +1$, (12) reduces to

$$\begin{array}{l} X_1 = \mu_1 + \sigma_1 Z_1 \\ X_2 = \mu_2 + \sigma_2 Z_1 \end{array} \implies \begin{array}{l} X_1 = \mu_1 + \sigma_1 Z_1 \\ X_2 = \mu_2 + \sigma_2\left(\frac{X_1-\mu_1}{\sigma_1}\right) = \mu_2 - \frac{\sigma_2}{\sigma_1}\mu_1 + \frac{\sigma_2}{\sigma_1}X_1, \end{array}$$

and we see that X_2 is just a linear function of X_1 with positive slope. Similarly, at the other extreme, when $\rho = -1$, we can show $X_2 = \mu_2 + \frac{\sigma_2}{\sigma_1}\mu_1 - \frac{\sigma_2}{\sigma_1}X_1$. Therefore, when $|\rho| = 1$ we have that one of these variables is a direct linear function of the other and, in these cases, the support of (X_1, X_2) will be a linear submanifold of $\mathbb{R}^2$ of dimension strictly less than 2; therefore, for this reason, we cannot expect there to exist a joint pdf.

Transformations like (12) can often be exploited to simplify computations involving bivariate normals.

Example.
Suppose (X_1, X_2) is bivariate normal with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$.
Derive the conditional distribution of X_2 given X_1.
Note: if we try to write down

$$f_{X_2|X_1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)},$$

then it is tedious to recognize what the mean and variance are for the underlying normal even though recognizing it will be normal is not difficult.

SOLUTION:
Using the transformation (12) we see X_1 is *independent* of Z_2 since X_1 is only a function of Z_1 (which is independent of Z_2). Therefore,

$$\begin{aligned} X_2 &= \mu_2 + \sigma_1\rho Z_1 + \sigma_2\sqrt{1-\rho^2}Z_2 \\ &= \underbrace{\mu_2 + \sigma_2\rho\left(\frac{X_1-\mu_1}{\sigma_1}\right)}_{\text{given } X_1\text{, this term is constant}} + \sigma_2\sqrt{1-\rho^2}Z_2. \end{aligned}$$

Therefore,

$$X_2|X_1 \sim \mathrm{N}\left(\mu_2 + \sigma_2\rho\left(\frac{X_1-\mu_1}{\sigma_1}\right), \sigma_2^2(1-\rho^2)\right).$$

Note that the first parameter is $E(X_2|X_1)$, and the second parameter is $Var(X_2|X_1)$.

Remark. (non-uniqueness of the linear transformation defining the bivariate normal)
There happens to be *many* linear transformations of Z_1 and Z_2 that lead to the *same* bivariate normal distribution. For example,

$$\begin{array}{l} X_1 = \mu_1 + \sigma_1\sqrt{1-\rho^2}Z_1 + \sigma_1\rho Z_2 \\ X_2 = \mu_2 \qquad\qquad\qquad + \ \sigma_2 Z_2 \end{array} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \sigma_1\sqrt{1-\rho^2} & \sigma_1\rho \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \tag{13}$$

is another linear transformation of Z_1 and Z_2 that leads to the same bivariate normal, i.e., having the exact same parameters. Please check that with this transformation $\mu_i = E(X_1),\ \sigma_i^2 = Var(X_i)$ and $\rho = \mathrm{corr}(X_1, X_2)$.

Alternate linear transformations are useful as hopefully the next exercise illustrates.

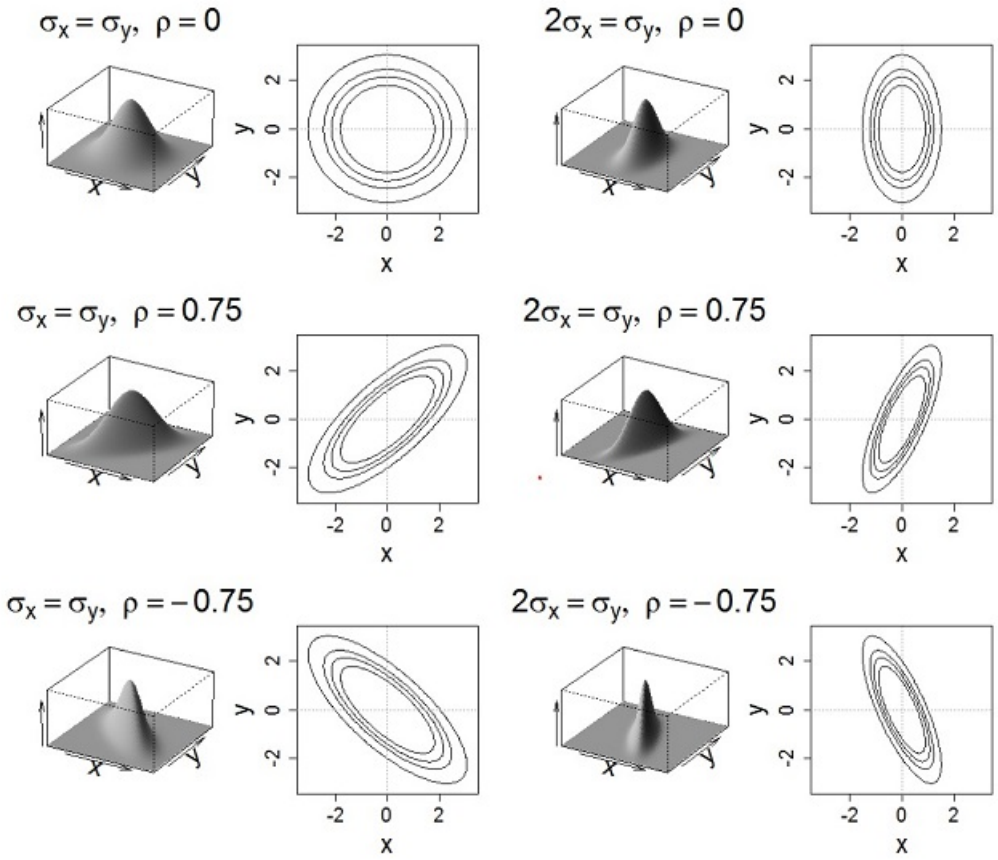
Exercise for the student.
Suppose (X_1, X_2) is bivariate normal with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$.
Derive the conditional distribution of X_1 given X_2, and use this to determine $E(X_1|X_2)$ and $Var(X_1|X_2)$.

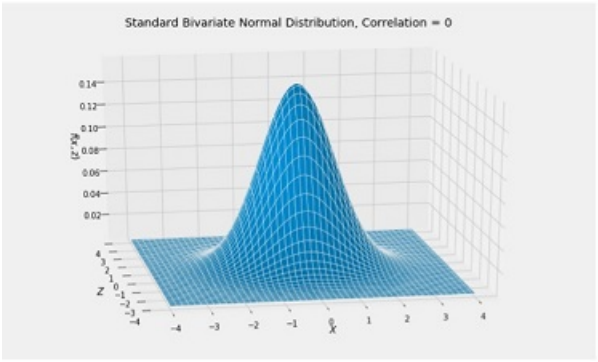
Note about this exercise:
Using the transformation (12) would not be ideal because doing so you'd have $Z_1 = \frac{X_2-\mu_2-\sigma_2\sqrt{1-\rho^2}Z_2}{\sigma_2\rho}$ and once given information about X_2 we'd have information not only about Z_1 but *also* Z_2, in particular, X_2 is *not* independent of Z_2. This is not like in the last example, where X_1 would not influence Z_2. Nevertheless, using the transformation (13) we will not run into this problem. The student should show that

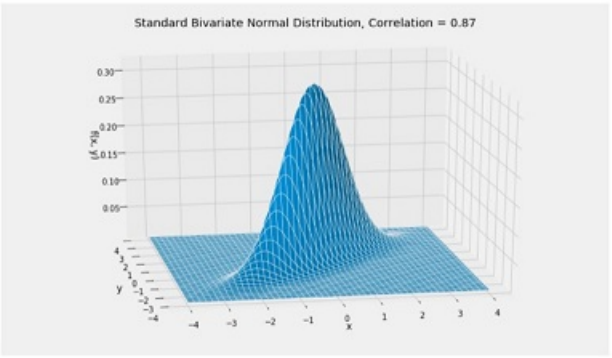
$$X_1|X_2 \sim \mathrm{N}\left(\mu_1 + \sigma_1\rho\left(\frac{X_2-\mu_2}{\sigma_2}\right), \sigma_1^2(1-\rho^2)\right),$$

and the parameters listed here, respectively, are $E(X_1|X_2)$ and $Var(X_1|X_2)$.

We end this section some pictures of the bivariate normal distribution.







The multivariate Normal distribution.*

This section requires the reader knows some basic linear algebra.

Suppose $k \geq 2$ is a fixed integer. In a way similar to how we defined the bivariate normal as a linear transformation of iid standard normals, we now define what it means for the vector $(X_1, X_2, \ldots, X_k)^T$ to have a k-variate normal distribution.

We say

$$\mathbf{X} := \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$$

has a **k-variate**, or, simply **multivariate normal distribution** with parameters $\vec{\mu}$ and $\mathbf{\Sigma}$, abbreviated $\mathbf{X} \sim \mathbf{N}_k(\vec{\mu}, \mathbf{\Sigma})$, provided, for some integer $\ell \geq 1$,

$$\mathbf{X} = \vec{\mu} + \mathbf{A}\mathbf{Z}, \tag{14}$$

where

$$\mathbf{Z} := \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_\ell \end{bmatrix}, \quad Z_1, Z_2, \ldots, Z_\ell \sim \text{iid N}(0,1), \qquad \vec{\mu} := \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \in \mathbb{R}^k$$

and $\mathbf{A}$ is a $k \times \ell$ real matrix with the property that

$$\mathbf{A}\mathbf{A}^T = \mathbf{\Sigma}.$$

Example. The bivariate normal is a 2-variate normal distribution.
Verify these facts for the case $k = 2$, i.e., the bivariate normal.

SOLUTION: From equation (12) on page 265 we see that $\mathbf{X} = \vec{\mu} + \mathbf{A}\mathbf{Z}$, where

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\sqrt{1-\rho^2} \end{bmatrix}, \quad \text{and} \quad \mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$$

with $Z_1, Z_2 \sim$ iid N$(0,1)$. Moreover,

$$\mathbf{A}\mathbf{A}^T = \begin{bmatrix} \sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} \sigma_1 & \sigma_2\rho \\ 0 & \sigma_2\sqrt{1-\rho^2} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

whose entries are exactly the covariances as required. Note that if $|\rho| = 1$, then $\mathbf{A}\mathbf{A}^T = \mathbf{\Sigma}$ is singular and, therefore, not positive definite, so no joint pdf of X_1, X_2 will exist in this case. □

Remark (**The parameters are $\vec{\mu} = E(\mathbf{X})$ and $\mathbf{\Sigma}$ is the matrix of covariances**). If we write

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1\ell} \\ a_{21} & a_{22} & \cdots & a_{2\ell} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{k\ell} \end{bmatrix},$$

then, working out the matrix multiplication, (14) can be alternately written as

$$\begin{aligned} X_1 &= \mu_1 + a_{11}Z_1 + a_{12}Z_2 + \cdots + a_{1\ell}Z_\ell \\ X_2 &= \mu_2 + a_{21}Z_1 + a_{22}Z_2 + \cdots + a_{2\ell}Z_\ell \\ &\vdots \\ X_k &= \mu_k + a_{k1}Z_1 + a_{k2}Z_2 + \cdots + a_{k\ell}Z_\ell. \end{aligned}$$

As each X_i is an affine transformation of independent normals, each X_i is normal; in fact, for each i,

$$X_i \sim \mathrm{N}(\mu_i, \sum_{m=1}^{\ell} a_{im}^2).$$

Therefore, $\vec{\mu}$ represents the vector of means or, simply, the ***mean vector***.

More generally,

$$\begin{aligned} Cov(X_i, X_j) &= E[(X_i - \mu_i)(X_j - \mu_j)] \\ &= E\left[\sum_{m=1}^{\ell} a_{im}Z_m \sum_{n=1}^{\ell} a_{jn}Z_n\right] \\ &= \sum_{m=1}^{\ell}\sum_{n=1}^{\ell} a_{im}a_{jn}E\left[Z_m Z_n\right] \\ &= \underbrace{\sum_{m=1}^{\ell}\sum_{n=1}^{\ell}}_{m=n} a_{im}a_{jm}\underbrace{E\left[Z_m^2\right]}_{=1} \quad + \quad \underbrace{\sum_{m=1}^{\ell}\sum_{n=1}^{\ell}}_{m\neq n} a_{im}a_{jn}\underbrace{E\left[Z_m Z_n\right]}_{=0} \\ &= \sum_{m=1}^{\ell} a_{im}a_{jm} = (\mathbf{A}\mathbf{A}^T)_{ij}. \end{aligned}$$

Therefore, since the $(i,j)^{\text{th}}$ entry of $\mathbf{A}\mathbf{A}^T$ is $Cov(X_i, X_j)$, it follows that $\mathbf{A}\mathbf{A}^T = \mathbf{\Sigma}$ represents the ***covariance matrix*** (or variance-covariance matrix).

The previous calculations could have been done succinctly using (14):

$$E(\mathbf{X}) = E(\vec{\mu} + \mathbf{AZ}) = \vec{\mu} + \mathbf{A}\underbrace{E(\mathbf{Z})}_{=\mathbf{0}} = \vec{\mu}$$

and

$$\begin{aligned} E[(\mathbf{X}-\vec{\mu})(\mathbf{X}-\vec{\mu})^T] &= E[\mathbf{AZ}(\mathbf{AZ})^T] \\ &= E[\mathbf{AZZ}^T\mathbf{A}^T] \\ &= \mathbf{A}\underbrace{E[\mathbf{ZZ}^T]}_{=I_k}\mathbf{A}^T = \mathbf{AA}^T, \end{aligned}$$

where I_k is the $k \times k$ identity matrix since the Z_i's are iid N(0,1). □

The k-variate normal density.
If $\mathbf{\Sigma}$ is ***positive definite***, then $\mathbf{X}$ has a joint pdf given as

$$f(\mathbf{x}) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\vec{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\vec{\mu})}}{(2\pi)^{\frac{k}{2}}\det(\mathbf{\Sigma})^{\frac{1}{2}}} \quad \text{for } \mathbf{x} \in \mathbb{R}^k.$$

This follows by method of Jacobians: for $\mathbf{z} \in \mathbb{R}^k$,

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}}}{(2\pi)^{\frac{k}{2}}}.$$

Since $\mathbf{\Sigma}$ is positive definite I claim that $\mathbf{A}$ can be taken to be $k \times k$ with $\det(\mathbf{A}) > 0$ and, thus, invertible: (***insert linear algebra result here***).

Now, $\mathbf{x} = \vec{\mu} + \mathbf{Az} \implies \mathbf{z} = \mathbf{A}^{-1}(\mathbf{x} - \vec{\mu})$. Moreover,

$$|J| = \left|\det\left(\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right)\right| = \left|\det(\mathbf{A}^{-1})\right| = \frac{1}{\det(\mathbf{A})}.$$

In fact, since $\det(\mathbf{\Sigma}) = \det(\mathbf{AA}^T) = \det(\mathbf{A})\det(\mathbf{A}^T) = \det(\mathbf{A})^2$, it follows

$$\det(\mathbf{A}) = \det(\mathbf{\Sigma})^{\frac{1}{2}}.$$

Therefore,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{\mathbf{Z}}(\mathbf{z}) \cdot |J| = \\ &= f_{\mathbf{Z}}(\mathbf{A}^{-1}(\mathbf{x}-\vec{\mu})) \cdot \frac{1}{\det(\mathbf{\Sigma})^{\frac{1}{2}}} \\ &= \frac{e^{-\frac{1}{2}(\mathbf{x}-\vec{\mu})^T(\mathbf{A}^{-1})^T\mathbf{A}^{-1}(\mathbf{x}-\vec{\mu})}}{(2\pi)^{\frac{k}{2}}\det(\mathbf{\Sigma})^{\frac{1}{2}}} = \frac{e^{-\frac{1}{2}(\mathbf{x}-\vec{\mu})^T(\mathbf{AA}^T)^{-1}(\mathbf{x}-\vec{\mu})}}{(2\pi)^{\frac{k}{2}}\det(\mathbf{\Sigma})^{\frac{1}{2}}} \\ &= \frac{e^{-\frac{1}{2}(\mathbf{x}-\vec{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\vec{\mu})}}{(2\pi)^{\frac{k}{2}}\det(\mathbf{\Sigma})^{\frac{1}{2}}}. \end{aligned}$$

□

VII. Inequalities and limit theorems.

Inequalities play a special role in probability. We start with two useful inequalities.

The Markov inequality.
Let X be a non-negative random variable. Then, for any constant $a > 0$,

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

Proof. We assume the random variable is continuous with pdf $f(x)$.

$$\begin{aligned} E(X) = \int_0^\infty x f(x)\,dx &= \underbrace{\int_0^a x f(x)\,dx}_{\text{this term} \geq 0} + \int_a^\infty x f(x)\,dx \\ &\geq 0 + \int_a^\infty x f(x)\,dx \\ &\overset{(*)}{\geq} \int_a^\infty a f(x)\,dx \\ &= a \int_a^\infty f(x)\,dx = a \cdot P(X \geq a). \end{aligned}$$

Therefore, we've shown $a \cdot P(X \geq a) \leq E(X)$ and, consequently, $P(X \geq a) \leq \frac{E(X)}{a}$. We point out that in $^{(*)}$ we used the fact that $xf(x) \geq af(x)$ for all $x \geq a$. □

Remark.
Since $X \geq 0$, $E(X)$ is either finite or infinitely positive ($E(X) = \infty$). Since $E(X) = \infty$ makes the inequality trivial, we usually restrict the use of this inequality to r.v.s with $E(X) < \infty$. In this case, the Markov inequality roughly says: if a nonnegative random variable has a small mean, then the probability it takes large values must also be small.

Remark.
The upper bound provided by Markov's inequality can be quite "loose".

Example.
Suppose $X \sim$ uniform(0, 4), then $E(X) = 2$ and:

Markov's inequality says...	whereas the actual values are...
$P(X \geq 2) \leq \frac{2}{2} = 1$	$P(X \geq 2) = 0.5$
$P(X \geq 3) \leq \frac{2}{3} = 0.6\bar{6}$	$P(X \geq 3) = 0.25$
$P(X \geq 4) \leq \frac{2}{4} = 0.5$	$P(X \geq 4) = 0$

As a corollary to the Markov inequality we have...

The Chebyshev inequality.
Let Y be any random variable having mean μ_Y and finite variance σ_Y^2. Then, for any constant $k > 0$,

$$P(|Y - \mu_Y| \geq k) \leq \frac{\sigma_Y^2}{k^2} \tag{15}$$

which is equivalent to

$$P(|Y - \mu_k| < k) \geq 1 - \frac{\sigma_Y^2}{k^2} \tag{16}$$

Remark.
This inequality roughly says: if a random variable has a small variance, then the probability it takes values far away from its mean is also small.

Remark.
If we replace k by $k\sigma_Y$ in form (15) and (16) of the Chebyshev's inequality, we get, respectively,

$$P(|Y - \mu_Y| \geq k\sigma_Y) \leq \frac{1}{k^2} \tag{17}$$

which is equivalent to

$$P(|Y - \mu_Y| < k\sigma_Y) \geq 1 - \frac{1}{k^2}. \tag{18}$$

Inequalities (15), (16), (17), and (18) are the different forms of Chebyshev inequalities. In the form (18), the inequality provides an interesting lower bound: this inequality says that the probability an rv takes values *within* k standard deviations from its mean is *at least* $1 - \frac{1}{k^2}$. So, in particular, there's a probability at least $\frac{3}{4}$ an rv takes values within 2 standard deviations of its mean and a probability at least $\frac{8}{9}$ it takes values within 3 standard deviations of its mean, etc.

Proof.
Let Y be any rv with finite mean μ_Y and finite variance σ_Y and let $k > 0$. Then

$$\begin{aligned} P(|Y - \mu_Y| \geq k) &= P(|Y - \mu_Y|^2 \geq k^2) \quad \text{since } (|Y - \mu_Y| \geq k) = (|Y - \mu_Y|^2 \geq k^2 \\ &\leq \frac{E(|Y - \mu_k|^2)}{k^2} \quad \text{by Markov's inequality} \\ &= \frac{\sigma_Y^2}{k^2}. \end{aligned}$$

An interesting application is the weak law of large numbers (WLLN).

The weak law of large numbers.
Suppose $X_1, X_2, X_3, \ldots$ are i.i.d. each with mean μ. Set $S_n := \sum_{i=1}^{n} X_i$ and let $\varepsilon > 0$ be arbitrary. Then

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) = 1.$$

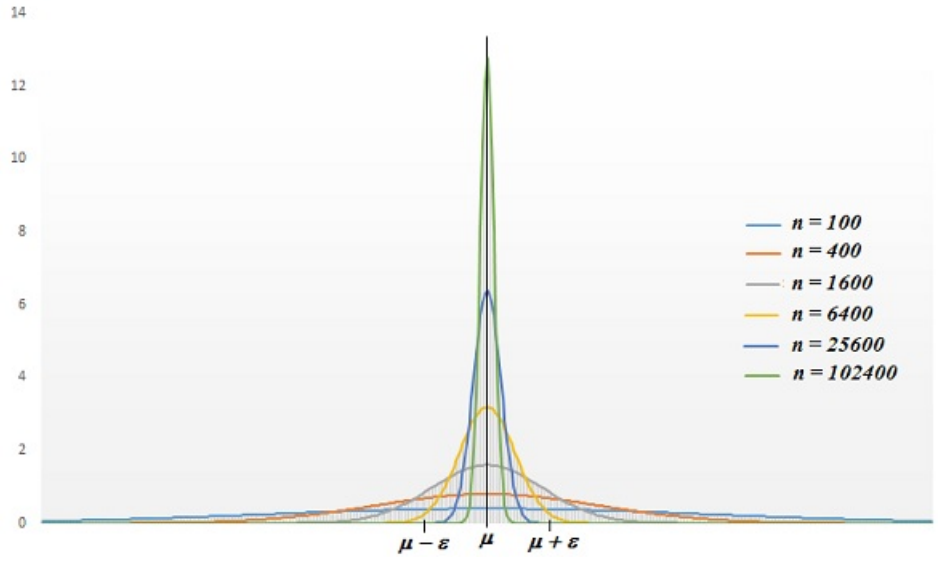
This theorem says: for any fixed tolerance $\epsilon > 0$, the probability that the sample mean is within ε of μ approaches 1 as the sample size n tends to $+\infty$.

Remark. (Paraphrasing the summarizing sentence above.)
Since S_n/n is really the sample mean of an iid sample $X_1, X_2, \ldots, X_n$, another way to think about the WLLN is as follows:

For any $\varepsilon > 0$ and $\delta > 0$, there exists an integer N such that for all $n \geq N$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) \geq 1 - \delta$$

for all $n \geq N$.



Proof.
For the proof we make the simplifying assumption that the identical distribution admits a finite variance σ^2 (but the result is true without it). For each n, let $S_n = \sum_{i=1}^{n} X_i$, where $X_1, X_2, \ldots$ is a sequence of iid rvs with mean μ and variance $\sigma^2 < \infty$. Note that

$$E\left(\frac{S_n}{n}\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu,$$

and

$$Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Now, fix an arbitrary $\varepsilon > 0$. By Chebyshev's inequality (16) using $k = \varepsilon, Y = \frac{S_n}{n}$,

$$P(|\frac{S_n}{n} - \mu| < \varepsilon) \geq 1 - \frac{Var\left(\frac{S_n}{n}\right)}{\varepsilon^2} = 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

Therefore,

$$1 - \frac{\sigma^2}{n\varepsilon^2} \leq P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) \leq 1.$$

Take the limit as n tends to $+\infty$ by Squeeze Theorem. □

Application: Monte-carlo method.*

The Monte-Carlo method is much more general, but we will demonstrate the idea with hopefully a simple example.

Example.
Estimate $\mathcal{I} := \int_0^1 g(u)\,du$ for some (possibly complicated) function $g(u)$. Assume g is bounded by a known constant M: $|g(u)| \le M$ for all $u \in [0,1]$.

SOLUTION:
Let $U_1, U_2, U_3, \ldots$ be iid uniform(0, 1). Let $X_i = g(U_i)$. Then $X_1, X_2, X_3, \ldots$ are also iid and

$$E(X_i) = E(g(U_i)) = \int_0^1 g(u)\cdot 1\,du = \int_0^1 g(u)\,du = \mathcal{I},$$

and

$$Var(X_i) \le E(g(U_i)^2) = \int_0^1 g(u)^2\,du \le M^2.$$

Therefore, the conditions of the weak law of large numbers are satisfied with

$$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^n g(U_i), \quad E\left(\frac{S_n}{n}\right) = \mathcal{I}, \quad \text{and} \quad Var\left(\frac{S_n}{n}\right) = \frac{Var\left(g(U_1)\right)}{n} \le \frac{M^2}{n}.$$

Therefore, for any fixed $\varepsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n g(U_i) - \int_0^1 g(u)\,du\right| < \varepsilon\right) \ge 1 - \frac{Var\left(\frac{S_n}{n}\right)}{\varepsilon^2} \ge 1 - \underbrace{\frac{M^2}{n\epsilon^2}}_{\delta}$$

for every n. Now, $\frac{M^2}{n\varepsilon^2} < \delta$ when $n \ge \frac{M^2}{\delta\varepsilon^2}$. This says if we specify a fix tolerance $\varepsilon > 0$ and a probability $\delta > 0$, then $P\left(\left|\frac{1}{n}\sum_{i=1}^n g(U_i) - \int_0^1 g(u)\,du\right| < \varepsilon\right) \ge 1 - \delta$ for any $n \ge \frac{M^2}{\delta\varepsilon^2}$. We then estimate $\mathcal{I}$ by

$$\frac{1}{n}\sum_{i=1}^n g(U_i).$$

Remark.
The last example says to estimate $\mathcal{I}$ just simulate n iid uniform(0,1)'s $U_1, U_2, \ldots, U_n$, plug each of them into the function g and then average the result! We'll be within ε of the actual value of $\mathcal{I}$ with a (very high) probability of at least $1 - \delta$.

Exercise for the student.
Take $g(u) = \frac{1}{1+u^2}$, $\varepsilon = \delta = .01$. Find an n so that $\frac{1}{n}\sum_{i=1}^n \frac{1}{1+U_i^2}$ will estimate $\mathcal{I} = \int_0^1 \frac{1}{1+u^2}\,du$ to within ε with probability at least .99. Then use a computer to estimate it. The actual value of $\mathcal{I}$ is $\frac{\pi}{4}$. How does it do?

The central limit theorem.

We showed that iid sequences possessing mean μ satisfy the Weak Law of Large Numbers[9]:

For every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \varepsilon\right) = 1.$$

Equivalently,[10]*, for* $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right) = 0.$$

We now discuss a refinement of the Weak Law...

The Central Limit Theorem (CLT).
Let $X_1, X_2, X_3, \ldots$ be an iid sequence of random variables having mean μ and *positive* but finite variance $0 < \sigma^2 < \infty$ (standard deviation σ). Let $S_n = \sum_{i=1}^{n} X_i$ for each n. Then,

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x).$$

Equivalently,

$$\lim_{n\to\infty} P\left(\frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) = \Phi(x).$$

We will prove this theorem under the additional (simplifying) assumption that the identical distribution possesses a moment generating function. For the proof we will need the following property about MGFs (stated without proof) as well as some reminders about the properties of the MGF and basic calculus. For students who do not want to see the proof, you may jump to page 281.

The Continuity Theorem of Moment-Generating Functions.
If we have a sequence (Y_n) of random variables that each have an MGF and $\lim_{n\to\infty} M_{Y_n}(\theta) = M_X(\theta)$ for all θ in an open interval containing 0, where $M_X(\theta)$ is a MGF of a rv X having a continuous CDF, then for all real x:

$$\lim_{n\to\infty} P(Y_n \leq x) = P(X \leq x).$$

i.e., the distribution of Y_n is converging to the distribution of X as n approaches infinity.

[9]Our proof of the WLLNs required the distribution possess a variance as well, but, in fact, this is not necessary.

[10]We usually say $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ in probability and write $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu$ as $n \to \infty$.

All MGFs have the following property:

Excercise for the student. (solution below... don't peek until you've attempted this!)
If a and b are any constants and X has MGF $M(\theta)$, then the MGF of $aX + b$ is

$$M_{aX+b}(\theta) = e^{\theta b} M(a\theta).$$

Corollary. (The MGF of a standardized rv)
If X has mean μ and standard deviation σ, then

$$M_{\frac{X-\mu}{\sigma}}(\theta) = e^{-\frac{\mu\theta}{\sigma}} M\left(\frac{\theta}{\sigma}\right).$$

Proof.
Take $a = \frac{1}{\theta}$ and $b = -\frac{\mu}{\theta}$ in the exercise.

Solution to Exercise.

$$\cdot(\upsilon\theta)\mathcal{M}_{q\theta^{\partial}} = ({}_{X(\upsilon\theta)^{\partial}})\mathcal{E} \cdot {}_{q\theta^{\partial}} = ({}_{q\theta^{\partial}} \cdot {}_{X\upsilon\theta^{\partial}})\mathcal{E} = ({}_{q\theta+X\upsilon\theta^{\partial}})\mathcal{E} = ({}_{[q+X\upsilon]\theta^{\partial}})\mathcal{E} = (\theta)^{q+X\upsilon}\mathcal{M}$$

We need to recall this calculus fact from page 106:

When c is a constant,

$$\lim_{n\to\infty}\left(1+\frac{c}{n}\right)^n = e^c.$$

In fact,

$$\lim_{n\to\infty}\left(1+\frac{c}{n}+o\left(\frac{1}{n}\right)\right)^n = e^c.$$

Note: for the "little oh" notation go to page 104 of these notes.

Finally, we'll also need the
The MacLaurin expansion of the MGF – see remark on page 149:

$$M_X(\theta) = 1 + E(X)\theta + E(X^2)\frac{\theta^2}{2!} + E(X^3)\frac{\theta^3}{3!} + \cdots.$$

Proof of CLT.
Suppose $X_1, X_2, \cdots \sim$ iid having MGF $M(\theta)$, $E[X_i] = \mu$ and $E[X_i^2] = \sigma^2 + \mu^2$ for all i.
Let's find the MGF of

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}}.$$

$$\begin{aligned}
M_{Y_n}(\theta) &= E(e^{\theta Y_n}) = E\left(e^{\theta\cdot\frac{\sum(X_i-\mu)}{\sigma\sqrt{\sigma\sqrt{n}}}}\right) = E\left(\prod_{i=1}^n e^{\frac{\theta}{\sigma\sqrt{n}}(X_i-\mu)}\right)\\
&= \prod_{i=1}^n E\left(e^{\frac{\theta}{\sigma\sqrt{n}}X_i}\cdot e^{-\frac{\mu\theta}{\sigma\sqrt{n}}}\right) \qquad \text{by independence}\\
&= \prod_{i=1}^n e^{-\frac{\mu\theta}{\sigma\sqrt{n}}}\cdot M\left(\frac{\theta}{\sigma\sqrt{n}}\right)\\
&= \left[e^{-\frac{\mu\theta}{\sigma\sqrt{n}}}\cdot M\left(\frac{\theta}{\sigma\sqrt{n}}\right)\right]^n
\end{aligned}$$

Let's look closer at the last term...

$$\begin{aligned}
&e^{-\frac{\mu\theta}{\sigma\sqrt{n}}}\cdot M\left(\frac{\theta}{\sigma\sqrt{n}}\right)\\
&= \left(1-\frac{\mu\theta}{\sigma\sqrt{n}}+\frac{1}{2!}\cdot\frac{(\mu\theta)^2}{\sigma^2 n}+o\left(\frac{1}{n}\right)\right)\cdot\left(1+\frac{\mu\theta}{\sigma\sqrt{n}}+\frac{1}{2!}\cdot\frac{(\sigma^2+\mu^2)\theta^2}{\sigma^2 n}+o\left(\frac{1}{n}\right)\right)\\
&= 1-\frac{\mu\theta}{\sigma\sqrt{n}}+\frac{1}{2!}\cdot\frac{\mu^2\theta^2}{\sigma^2 n}+\frac{\mu\theta}{\sigma\sqrt{n}}-\frac{\mu^2\theta^2}{\sigma^2 n}+\frac{1}{2!}\cdot\frac{(\sigma^2+\mu^2)\theta^2}{\sigma^2 n}+o\left(\frac{1}{n}\right)\\
&= 1+\frac{\theta^2}{2n}+o\left(\frac{1}{n}\right).
\end{aligned}$$

Substituting this back...

$$\begin{aligned} M_{Y_n}(\theta) &= \left[1 + \frac{\theta^2}{2n} + o\left(\frac{1}{n}\right)\right]^n \\ &= \left[1 + \frac{\frac{\theta^2}{2}}{n} + o\left(\frac{1}{n}\right)\right]^n \to \quad e^{\frac{\theta^2}{2}} = M_Z(\theta) \text{ as } n \to \infty. \end{aligned}$$

Therefore, the result follows by the continuity theorem of MGFs. □

Intuitive meaning of the Central Limit Theorem.
For large n,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \approx Z \sim \mathrm{N}(0,1),$$

where $\approx$ means that the distribution of the rv on the left is approximately the distribution of a standard normal, and the larger the n the better the approximation. (more on this later). Equivalently,

$$S_n \approx \mathrm{N}(n\mu, n\sigma^2).$$

Therefore, if we can recognize a random variable S_n as a large sum of iid rvs with finite mean and finite positive variance, then by simply replacing the random variable S_n by a $\mathrm{N}(n\mu, n\sigma^2)$ rv we should get a reasonable approximation to the desired probability.

Example.
Suppose $U_1, U_2, \ldots, U_{1000} \sim$ iid unif$(0,1)$. Estimate $P(480 \le U_1 + U_2 + \cdots + U_{1000} \le 520)$.

SOLUTION: Let $S = U_1 + U_2 + \cdots + U_{1000}$. Since $E(U_i) = \frac{1}{2}$ and $Var(U_i) = \frac{1}{12}$,

$$E(S) = 1000 \cdot \tfrac{1}{2} = 500 \quad \text{and} \quad Var(S) = 1000 \cdot \tfrac{1}{12} \ (\sigma_S \approx 9.1287).$$

Therefore, assuming $n = 1000$ is large, $S \approx \mathrm{N}(500, \frac{1000}{12})$ and

$$\begin{aligned} P(480 \le S \le 520) &\approx P\left(480 \le Y \le 520\right). \\ &= P\left(\frac{480-500}{9.1287} \le \frac{Y-500}{\sqrt{\frac{1000}{12}}} \le \frac{520-500}{9.1287}\right) \\ &= \Phi\left(\frac{520-500}{9.1287}\right) - \Phi\left(\frac{480-500}{9.1287}\right) \\ &= 0.9857 - 0.0143 = 0.9714. \end{aligned}$$

Exercise for the student.
Suppose $U_1, U_2, U_3 \sim$ iid unif$(0,1)$. Use the CLT to estimate $P(U_1 + U_2 + U_3 < \frac{1}{2})$. Try to compute this probability exactly. How did your approximation do?
ANSWER: CLT : $P(S < \frac{1}{2}) \approx .0228$, Actual: $\int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}-x} \int_0^{\frac{1}{2}-x-y} 1 \, dzdydx = \frac{1}{48} \approx .0208$.
Seems good!

Following up with the last example...

Example.
Estimate the probability that you need at least 49 uniform(0,1)'s to see S_n exceed 28.

SOLUTION:
The event of interest says that $U_1+\cdots+U_{48} \leq 28$, which is equivalent to saying we'll need at least 49 observations to exceed 28. Since $E(S) = 48 \cdot \frac{1}{2} = 24$, $Var(S) = 48 \cdot \frac{1}{12} = 4$, and $\sigma_S = 2$.

$$\begin{aligned} P(U_1 + U_2 + \cdots + U_{48} \leq 28) &= P(S \leq 28) \\ &= P\left(\frac{S - \mu_S}{\sigma_S} \leq \frac{28-24}{2}\right) \\ &\overset{\text{CLT}}{\approx} \Phi(2) = 0.9772. \end{aligned}$$

□

Example.(The Chi-squared distribution with n degrees of freedom)
Let $n \geq 1$ be an integer. Suppose $X \sim X_n^2$, i.e. X has a Chi-squared distribution with n degrees of freedom. One can define this as the distribution of

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2,$$

where $Z_1, Z_2, \ldots, Z_n \sim$ iid Normal(0,1). The student should show (exercise) that $X \sim$ Gamma$\left(\frac{n}{2}, 2\right)$ using MGFs. Therefore,

$$P\left(\frac{X-n}{\sqrt{2n}} \leq x\right) = P(X \leq n + x\sqrt{2n}) \to \Phi(x) \text{ as } n \to \infty.$$

Note that this is a limiting result (as $n \to \infty$). Practically, we'd expect for large *fixed* n that the value of the probability on the left will be "close" to $\Phi(x)$.

The last examples had the identical distribution being continuous. However, the CLT also works when the identical distribution is discrete.

Example. (**Normal approximation to the binomial**)
If $X_1, X_2, X_3, \ldots$ are iid Bernoulli(p) with $0 < p < 1$, then $S_n = \sum_{i=1}^n X_i \sim \text{binom}(n, p)$. Moreover, since $E(S_n) = np$ and $Var(S_n) = np(1-p)$ exist and are finite, the CLT guarantees for every $u \in \mathbb{R}$,

$$P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq u\right) \to \Phi(u) \text{ as } n \to \infty.$$

This means that for sufficiently large n the two CDFs will be as close as we'd like[11].

Let $S_n \sim \text{binom}(n, p)$ with fixed p. From footnote[11], if n is such that $np(1-p) \geq 5$ then

$$\frac{S_n - np}{\sqrt{np(1-p)}} \approx Z \quad \Longrightarrow \quad S_n \approx n\mu + \sqrt{np(1-p)}Z.$$

As a concrete illustration suppose $S_{20} \sim \text{binom}(20, 0.5)$. Then, since $20(.5)(1-.5) = 5$, the CLT in this case with $n = 20$ should give a "decent" approximation. We can try to use the CLT to estimate $P(13 \leq S_{20} \leq 15)$. Now, $E(S_{20}) = 10$, $Var(S_{20}) = 5$, and, therefore, $S_{20} \approx 10 + \sqrt{5}Z$, equivalently, $\frac{S_{20}-10}{\sqrt{5}} \approx Z \sim \text{N}(0, 1)$:

$$\begin{aligned} P(13 \leq S_{20} \leq 15) &= P\left(\frac{13-10}{\sqrt{5}} \leq \frac{S_{20}-10}{\sqrt{5}} \leq \frac{15-10}{\sqrt{5}}\right) \\ &\approx P(1.34 \leq Z \leq 2.24) \\ &= \Phi(2.24) - \Phi(1.34) = 0.9875 - 0.9099 = 0.0776. \end{aligned}$$

Remark.
We should note that, in the last example, we can compute $P(13 \leq S_{20} \leq 15)$ *exactly*:

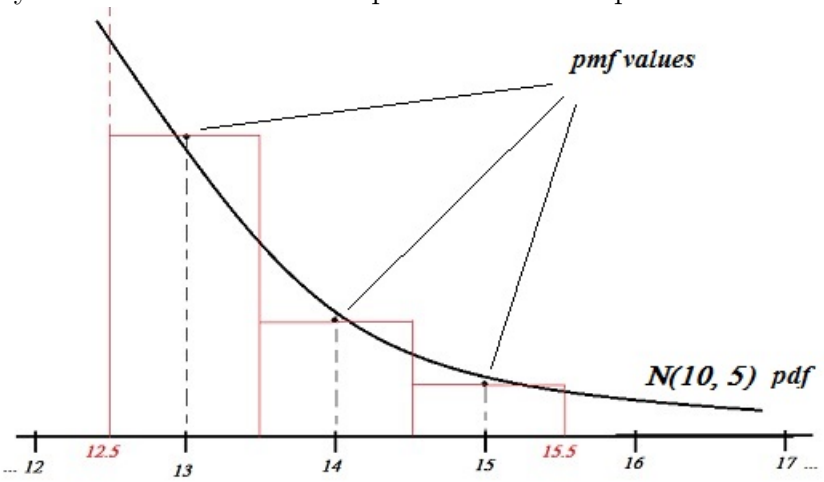
$$P(13 \leq S_{20} \leq 15) = \binom{20}{13}\left(\frac{1}{2}\right)^{20} + \binom{20}{14}\left(\frac{1}{2}\right)^{20} + \binom{20}{15}\left(\frac{1}{2}\right)^{20} = 0.125679\ldots,$$

and it appears the CLT approximation is not so sharp. Nevertheless, when dealing with binomial random variables (in fact, *any* discrete *integer-valued* rv) we may be able to improve upon the result of the CLT by using a so-called *continuity correction*.

[11] In the case of the binom(n, p) with p fixed, if $np(1-p) \geq 5$, then the CLT gives decent approximations, and the larger $np(1-p)$ is from 5, then better the resulting approximation.

Continuity corrections in using CLT for integer-valued random variables.
When the identical distribution is supported by integer values, the result of the CLT can often be improved by appropriately adding/subtracting 0.5 before invoking the CLT. We now explain the idea.

The issue that arises in using the CLT for approximating probabilities involving discrete rvs is that the exact probability is computed by *adding probability masses* in the event whereas the CLT approximation would be *integrating probability density* – i.e., finding area (under a pdf) – and these two operations are inherently *different*! Nevertheless, when the discrete random variable happens to be integer-valued, it turns out we can reconcile this difference to a great degree because, in this case, adding probability masses can be "nicely" converted into an area problem. See the picture below.



From the CLT, $S_{20} \approx \mathrm{N}(10, 5)$ but the $\mathrm{N}(10, 5)$ pdf will not perfectly interpolate the actual pmf values. The actual value of $P(13 \le S_{20} \le 15)$ is the sum of the 3 masses/heights at 13, 14, and 15. These heights are typically *unknown* but are being approximated by the pdf. The actual probability can now be interpreted as the area within the 3 red rectangles of width 1 centered at the values we are interested in: since these rectangles have width 1, each rectangle has area exactly equal to the pmf value there. From the picture it now seems a bit clearer that area under the pdf between 12.5 and 15.5 does a better job at estimating the area of these 3 red rectangles than had we found the area under the pdf between 13 and 15. This is the essence of the idea behind the ***0.5 continuity correction***.

Continuing with the last example, we try to improve upon the naïve CLT by making a .5 continuity correction before invoking the CLT.

Example (continued).
Since S_{20} is an integer-valued rv,

$$\begin{aligned} P(13 \le S_{20} \le 15) &= P(12.5 \le S_{20} \le 15.5) \\ &= P\left(\frac{12.5-10}{\sqrt{5}} \le \frac{S_{20}-10}{\sqrt{5}} \le \frac{15.5-10}{\sqrt{5}}\right) \\ &\overset{\text{CLT}}{\approx} P(1.11 \le Z \le 2.46) = \Phi(2.46) - \Phi(1.11) = .1266, \end{aligned}$$

which compares quite favorably to the actual value of .125679....

Example.
Suppose $S \sim \text{binom}(100, .2)$. Use the CLT to estimate $P(S = 20)$.

SOLUTION: First note that $Var(S) = 100(.2)(.8) = 16 \geq 5$ so the CLT should provide a decent approximation. In fact, $S \approx \text{N}(20, 16)$. Naïvely, without a continuity correction, $P(S = 20) \approx P(\frac{S-20}{4} = 0) \approx P(Z = 0) = 0$ since Z is a continuous rv whereas the actual value is $P(S = 20) = \binom{100}{20}.2^{20}.8^{80} = .0993\ldots$.

A continuity correction can offer a great improvement:

$$\begin{aligned} P(S = 20) &= P(19.5 \leq S \leq 20.5) \\ &= P(\frac{19.5 - 20}{4} \leq \frac{S - 20}{4} \leq \frac{20.5 - 20}{4}) \\ &\overset{\text{CLT}}{\approx} (-.125 \leq Z \leq .125) = .0994. \end{aligned}$$

□

Example.
If $X_n \sim \text{Poisson}(n)$, then we can think of X_n as a sum of iid Poisson(1) rvs:

$$X_n = Y_1 + Y_2 + \cdots + Y_n,$$

where $Y_1, Y_2, \cdots \sim$ iid Poisson(1).

$$E(X_n) = n, Var(X_n) = n \implies \sigma_{X_n} = \sqrt{n}.$$

With $n = 100$, say, the CLT gives

$$\begin{aligned} P(X_{100} \leq 110) &= P(X_{100} \leq 110.5) \\ &= P\left(\frac{X_{100} - \mu_{X_{100}}}{\sigma_{X_{100}}} \leq \underbrace{\frac{110.5 - 100}{10}}_{1.05} \right) \\ &\overset{\text{CLT}}{\approx} \Phi(1.05) = 0.8531. \end{aligned}$$

Note: we employed a continuity correction because the Poisson(100) is an integer-valued rv. In fact, the exact value can be found in this case (up to rounding error) to be $0.852862652\ldots$, and it appears the CLT did a good job here.

Example.
Estimate the probability we need at least 50 rolls of a fair die to see the sum exceed 180.

SOLUTION: Let $X_1, X_2, \ldots, X_n \sim$ iid discrete uniform$\{1,2,3,4,5,6\}$.
Set $S_n = X_1 + X_2 + \cdots + X_n$. Since $E(X_i) = \frac{7}{2} = 3.5$, $Var(X_i) = \frac{35}{12} = 2.916\overline{6}$, $\sigma_{X_i} = 1.707825\ldots$, we have

$$\mu_S = 49 \cdot \frac{7}{2} = 171.5, \quad \sigma_S^2 = 49 \cdot \frac{35}{12} = 142.916\ldots, \quad \text{and} \quad \sigma_S = \sqrt{142.916} \approx 11.955.$$

The event that 50 (or more) rolls are needed to see the sum exceed 180 is the same as saying that the sum S_{49} is less than or equal to 180. Therefore, we want to estimate $P(S_{49} \leq 180)$. Using a continuity correction (since S_{49} is discrete integer-valued):

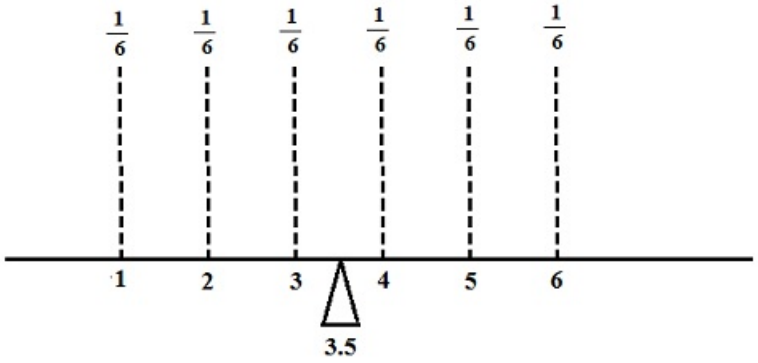
$$\begin{aligned} P(S_{49} \leq 180) &= P(S_{49} \leq 180.5) \\ &= P\left(\frac{S - \mu_S}{\sigma_S} \leq \frac{180.5 - 171.5}{11.955}\right) \overset{\text{CLT}}{\approx} \Phi(0.75) = 0.7734. \end{aligned}$$

Example.
Suppose now we only roll the die 3 times. Use the CLT to estimate the probability the total sum is 9. Compare your answer with the actual probability.

SOLUTION: $E(S_3) = 3 \cdot \frac{7}{2} = 10.5$, $Var(S_3) = 3 \cdot \frac{35}{12} = 8.75$ and $\sigma_{S_3} = 2.958$.

$$\begin{aligned} P(S_3 = 9) &= P(8.5 \leq S_3 \leq 9.5) = P\left(\frac{8.5 - 10.5}{2.958} \leq \frac{S_3 - 10.5}{2.958} \leq \frac{9.5 - 10.5}{2.958}\right) \\ &\overset{\text{CLT}}{\approx} P(-.68 \leq Z \leq -.34) = .1186. \end{aligned}$$

The actual probability is $P(S_3 = 9) = \frac{25}{216} = .1157\ldots$ (details omitted). In this case (even with n as small as 3) the CLT did a pretty good job with the estimation! This begs the question: can we always expect the CLT to give such a wonderful approximation for n as small as 3? The answer, of course, is *NO*! We will discuss this next, but state here that the reason this approximation worked so well in this example is because the identical distribution happens to be symmetric about its mean.



For fixed constants a and b we ask:

$$\textit{How large does } n \textit{ need to be in order for } \Phi(b) - \Phi(a) \textit{ to be "close" to } P\left(a < \tfrac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right)?$$

The answer depends on how symmetric the population distribution is about its mean. We present the following rules of thumb.

Heuristic.

- If the population is symmetric about its mean, the n as small as 10 gives good results (or, as the last example shows, n as small as 3).
- If the population is only mildly skewed (or if there is no reason to believe the population is heavily skewed) then $n \geq 30$ usually suffices for the CLT to give decent results, however,...
- The more skewed the population distribution, the larger n will need to be. For instance, if we have a Bernoulli(p) population, so that $S_n \sim \text{binom}(n, p)$, then (see footnote[11] on page 283)

  $$n \geq \frac{5}{p(1-p)}$$

  will give a reasonable approximation. Notice, in the case of the binomial, the closer p is to 0 (or to 1) the larger n will need to be; for example, if $p = 10^{-6}$ then (by this heuristic) n would need to be over 5 million for the CLT to give a "decent" approximation.

The following result is usually of more theoretical value than of practical value. It states the precise rate of convergence of the CDF to Φ.

The Berry-Essen Theorem. (simplified version to fit our statement of the CLT)
There exists a universal constant C such that:
if $X_1, X_2, \ldots$ are iid mean μ, variance σ^2, and $E(|X_i - \mu|^3) = \tau$, then with $S_n = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$,

$$E(S_n) = n\mu, Var(S_n) = n\sigma^2, \text{ and}$$

$$\left| P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) - \Phi(x) \right| \leq \frac{C \cdot \tau}{\sigma^3\sqrt{n}} = \frac{C \cdot E\left(\left|\frac{X_i - \mu}{\sigma}\right|^3\right)}{\sqrt{n}}$$

for all x and n.

The following interesting application doesn't *directly* follow from the statement of the central limit theorem the way we've stated it and I'll mention why.

Application: Stirling's approximation (see page 105)
For large integers n,

$$n! = \sqrt{2\pi n}\, n^n e^{-n}(1 + o(1)) \quad \text{as } n \to \infty,$$

which means that the relative error in approximating $n!$ by $\sqrt{2\pi n} n^n e^{-n}$, namely,

$$\frac{n! - \sqrt{2\pi n}\, n^n e^{-n}}{\sqrt{2\pi n}\, n^n e^{-n}}$$

goes to 0 as n tends to infinity.

Let's consider $P(X_n = n)$, where $X_n \sim \text{Poisson}(n)$.

On one hand,

$$P(X_n = n) = e^{-n} \cdot \frac{n^n}{n!}, \tag{19}$$

since we just need to evaluate the Poisson(n) pmf at the value n.

On the other hand, by the CLT, $\frac{X_n - n}{\sqrt{n}} \approx Z$ and employing a continuity correction (since X_n is integer-valued):

$$\begin{aligned} P(X_n = n) &= P(n - \frac{1}{2} \le X_n \le n + \frac{1}{2}) = P(-\frac{1}{2\sqrt{n}} \le \frac{X_n - n}{\sqrt{n}} \le \frac{1}{2\sqrt{n}}) \\ &\overset{\text{CLT}^*}{\approx} \Phi(\frac{1}{2\sqrt{n}}) = \Phi(-\frac{1}{2\sqrt{n}}) = \int_{-\frac{1}{2\sqrt{n}}}^{\frac{1}{2\sqrt{n}}} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz \\ &\approx \frac{e^0}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{2\pi n}}, \end{aligned}$$

i.e., by the CLT*,

$$P(X_n = n) \approx \frac{1}{\sqrt{2\pi n}}. \tag{20}$$

Therefore, from expressions (19) and (20) we have

$$e^{-n} \cdot \frac{n^n}{n!} = P(X_n = n) \approx \frac{1}{\sqrt{2\pi n}},$$

which simplifies to

$$n! \approx \sqrt{2\pi n}\, n^n e^{-n}.$$

□

* The issue with applying the CLT here is that, strictly speaking, the endpoints need to be fixed constants a and/or b, something like this: $a \le \frac{X_n - n}{\sqrt{n}} \le b$, so that the endpoints remain fixed as $n \to \infty$. However, in this application, the endpoints are functions of n; indeed, we have $-\frac{1}{2\sqrt{n}} \le \frac{X_n - n}{\sqrt{n}} \le \frac{1}{2\sqrt{n}}$ and this interval is collapsing to a single point as $n \to \infty$! Therefore, the CLT, by the way it has been developed, doesn't directly apply here. What is needed to make what we did rigorous is something called the ***local central limit theorem*** which we will not state.

The CLT implies the WLLN.
Suppose the $X_1, X_2, X_3, \ldots$ satisfy the conditions of the central limit theorem, and set $S_n = \sum_{i=1}^n X_i$. Then

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) &= P(-n\varepsilon < S_n - n\mu < n\varepsilon) \\ &= P\left(\frac{-n\varepsilon}{\sigma\sqrt{n}} < \frac{S_n - n\mu}{\sigma\sqrt{n}} < \frac{n\varepsilon}{\sigma\sqrt{n}}\right) \\ &= P\left(\frac{-\varepsilon\sqrt{n}}{\sigma} < \frac{S_n - n\mu}{\sigma\sqrt{n}} < \frac{\varepsilon\sqrt{n}}{\sigma}\right). \end{aligned}$$

Notice that, for any fixed $M > 0$,

$$\left(-M < \frac{S_n - n\mu}{\sigma\sqrt{n}} < M\right) \subseteq \left(\frac{-\varepsilon\sqrt{n}}{\sigma} < \frac{S_n - n\mu}{\sigma\sqrt{n}} < \frac{\varepsilon\sqrt{n}}{\sigma}\right)$$

when n is sufficiently large; and, in this case,

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) &\geq P\left(-M < \frac{S_n - n\mu}{\sigma\sqrt{n}} < M\right) \\ &\xrightarrow{\text{CLT}} \Phi(M) - \Phi(-M) \text{ as } n \to \infty \\ &\to 1 \text{ as } M \to \infty \end{aligned}$$

and, we see the CLT implies the WLLN!

Convex functions.
Let $-\infty \leq a < b \leq +\infty$. A function $h : (a, b) \to \mathbb{R}$ is called ***convex*** if for every $x, y \in (a, b)$,

$$h(\lambda x + (1 - \lambda)y) \leq \lambda h(x) + (1 - \lambda)h(y)$$

for all $0 \leq \lambda \leq 1$. In other words, a function h is convex if the line segment connecting $h(x)$ and $h(y)$ lies above the graph of the function there (see figure below).

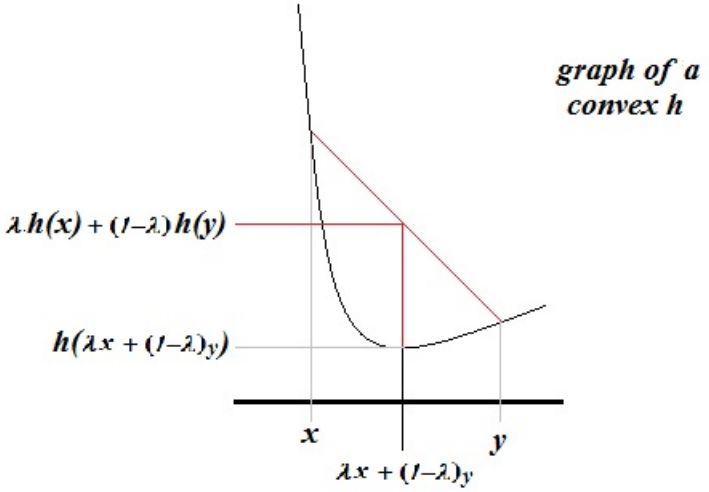
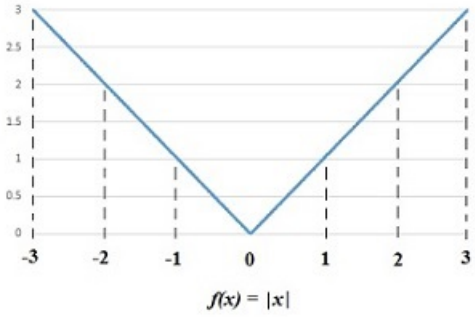
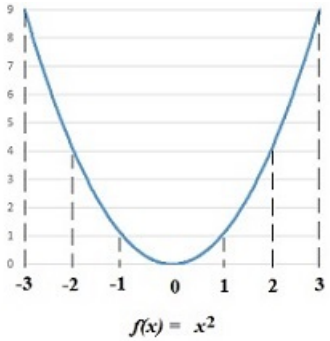
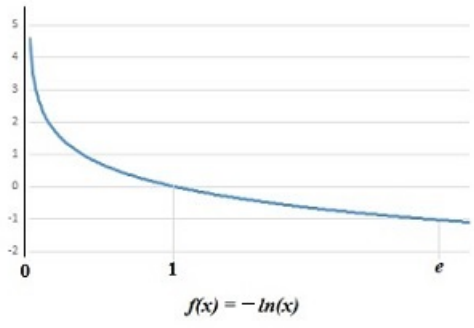


Figure. Graph of generic convex function: for every choice of x and y, the graph of the line segment connecting $h(x)$ and $h(y)$ *lies above* the graph of the function itself.

Here are pictures of some common convex functions defined on subintervals of $\mathbb{R}$:







Jensen's Inequality.
Let X be a random variable and $h(x)$ a convex function (defined on the support of X). Then

$$h(E(X)) \leq E(h(X)).$$

We will not present the proof of Jensen's inequality, but, instead, we will illustrate how it follows immediately from the definition of a convex function in the case of random variables taking only two values. For example, suppose X takes only the two values x_1 and x_2, where

$$P(X = x_1) = p \quad \text{and} \quad P(X = x_2) = 1 - p.$$

Notice that $E(X) = px_1 + (1-p)x_2$ is a convex combination of x_1 and x_2. Now, since h is convex, we have

$$h(E(X)) = h(px_1 + (1-p)x_2) \leq ph(x_1) + (1-p)h(x_2) = E[h(X)]$$

by the law of the unconscious statistician. Therefore, Jensen's inequality follows directly from the definition of a convex function in the case of random variables taking only two values.

Example. Finite second moment implies finite first moment.
Let X be any random variable with finite second moment. Since $h : \mathbb{R} \to \mathbb{R}$ defined by $h(x) = x^2$ is convex, Jensen's inequality says

$$[E(X)]^2 =: h(E(X)) \quad \leq \quad E[h(X)] := E(X^2).$$

Therefore, if $E(X^2) < \infty$, we must also have $[E(X)]^2 < \infty$, which implies $E(X) < \infty$.

Not only does a finite second moment imply a finite first moment, but in fact, a *much* more general statement is true:

Liapounov's inequality.
Let $0 < p < q$ be real numbers and let X be a random variable. Then

$$[E(|X|^p)]^{1/p} \leq [E(|X|^q)]^{1/q}.$$

That is, a finite qth moment implies a finite pth moment for any $0 < p < q$.

Proof. Left as an **exercise for the student.** Hint: $h(x) = |x|^{q/p}$ is a convex function. Now apply Jensen's inequality using this h and the random variable $|X|^p$.

The strong law of large numbers*

Let $X_1, X_2, X_3, \ldots$ be an iid sequence having finite mean $E(X_i) = \mu$. The ***strong law of large numbers*** says

$$P\left(\lim_{n\to\infty} \frac{\sum_{i=1}^n X_i}{n} = \mu\right) = 1, \tag{21}$$

or, equivalently, as

$$P\left(\lim_{n\to\infty} \frac{\sum_{i=1}^n X_i}{n} \neq \mu\right) = 0. \tag{22}$$

In words, this result says that, except for an event $N \subseteq \Omega$ of probability zero, if $\omega \in N^c$, then the *infinite sequence* $X_1(\omega), X_2(\omega), X_3(\omega), \ldots$ will have the property

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n X_i(\omega) = \mu.$$

It is certain that when we observe an infinite sequence of iid rvs that the running average of their values will converge to the mean μ, the set of sequences that don't have this property has probability zero.

Remark.
Please understand that the event we are dealing with in the strong law of large numbers involves an experiment, i.e., a sample space, that is *infinite-dimensional*!

Exercise for the student.
Show that the strong law implies the weak law. The following equivalent forms of the strong law might help:

$$P\left(\lim_{n\to\infty} \left|\frac{\sum_{i=1}^n X_i}{n} - \mu\right| = 0\right) = 1, \tag{23}$$

$$P\left(\lim_{n\to\infty} \left|\frac{\sum_{i=1}^n X_i}{n} - \mu\right| > 0\right) = 0. \tag{24}$$

Proof. We will prove the SLLN under the additional assumption $E(X_i^4) < \infty$.
We first prove the result when $E(X_i) = 0$ (i.e. when $\mu = 0$). Set $S_n = \sum_{i=1}^n X_i$. We want to show

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = 0\right) = 1.$$

The strategy of the proof is to show that the random variable $\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4$ has a finite expected value, i.e.,

$$E\left(\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4\right) < \infty. \tag{25}$$

From this it would have to follow that $\left(\frac{S_n}{n}\right)^4$ must be finite with probability 1, i.e.,

$$P\left(\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4 < \infty\right) = 1;$$

otherwise, $\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4$ would be infinite with positive probability and thus cannot have a finite expectation. Now, since convergent series must have terms that go to zero, it would follow that

$$P\left(\lim_{n\to\infty}\left(\frac{S_n}{n}\right)^4 = 0\right)$$

implying

$$P\left(\lim_{n\to\infty}\frac{S_n}{n} = 0\right)$$

giving our result.

So, we proceed to verify equation (25): since

$$\left(\frac{S_n}{n}\right)^4 \geq 0,$$

we can swap the expected value with the infinite series: (this is the Fubini-Tonelli theorem)

$$E\left(\sum_{n=1}^{\infty}\frac{S_n^4}{n^4}\right) = \sum_{n=1}^{\infty}\left(\frac{E\left(S_n^4\right)}{n^4}\right).$$

Now we analyze

$$S_n^4 = \left(\sum_{i=1}^{n}X_i\right)\left(\sum_{j=1}^{n}X_j\right)\left(\sum_{k=1}^{n}X_k\right)\left(\sum_{l=1}^{n}X_l\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}X_iX_jX_kX_l.$$

This sum will have terms like

$$\begin{array}{cccc} X_1X_1X_1X_1 & X_1X_1X_2X_1 & \dots & X_nX_nX_nX_1 \\ X_1X_1X_1X_2 & X_1X_1X_2X_2 & \dots & X_nX_nX_nX_2 \\ X_1X_1X_1X_3 & X_1X_1X_2X_3 & \dots & X_nX_nX_nX_3 \\ \vdots & \vdots & \ddots & \vdots \\ X_1X_1X_1X_n & X_1X_1X_2X_n & \dots & X_nX_nX_nX_n \end{array}$$

which is in one-to-one correspondence with the order of the indices

$$\begin{array}{cccc} (1,1,1,1) & (1,1,2,1) & \dots & (n,n,n,1) \\ (1,1,1,2) & (1,1,2,2) & \dots & (n,n,n,2) \\ (1,1,1,3) & (1,1,2,3) & \dots & (n,n,n,3) \\ \vdots & \vdots & \ddots & \vdots \\ (1,1,1,n) & (1,1,2,n) & \dots & (n,n,n,n) \end{array}$$

which is the sample space in rolling an "n-sided die" 4 times.

Question: How many 4-of-a-kinds are there?
Answer: There are n, namely, (i,i,i,i) for $i = 1, 2, \ldots, n$, and $E(X_i^4) = c_4$.

Question: How many 2 pairs are there?
Answer: There are $\binom{n}{2}\binom{4}{2} = 3n(n-1)$, and for $I \neq j$, $E(X_i^2 X_j^2) = \underbrace{E(X_i^2)}_{c_2} E(X_j^2) = c_2^2$.

Question: How many 3-of-a-kinds are there?
Answer: There are $\binom{4}{3}n(n-1) = 4n(n-1)$, and for $i \neq j$, $E(X_i^3 X_j) = E(X_i^3)\underbrace{E(X_j)}_{=0} = 0$.

Question: How many exactly one pair with the other two distinct are there?
Answer: There are $\binom{4}{2}n(n-1)(n-2)$ and for distinct i, j, k, $E(X_i^2 X_j X_k) = E(X_i^2)E(X_j)E(X_k) = 0$.

Question: How many have all 4 rolls distinct?
Answer: There are $n(n-1)(n-2)(n-3)$ and for distinct i, j, k, l, $E(X_i X_j X_k X_l) = E(X_i)E(X_j)E(X_k)E(X_l) = 0$.

$$
\begin{aligned}
E(S_n^4) &= E\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} X_i X_j X_k X_l\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} E(X_i X_j X_k X_l) \\
&= nE(X_1^4) \\
&\quad +3n(n-1)E(X_1^2)E(X_2^2) \\
&\quad +\underbrace{\binom{4}{3}n(n-1)E(X_1^3)E(X_2)}_{=0} \\
&\quad +\underbrace{\binom{4}{2}n(n-1)(n-2)E(X_1^2)E(X_2)E(X_3)}_{=0} \\
&\quad +\underbrace{n(n-1)(n-2)(n-3)E(X_1)E(X_2)E(X_3)E(X_4)}_{=0} \\
&= c_4 n + 3c_2^2 n(n-1).
\end{aligned}
$$

Thus,

$$\begin{aligned}\frac{E(S_n^4)}{n^4} &= \frac{c_4 n + 3c_2^2 n(n-1)}{n^4} \\ &= c_4 \cdot \frac{1}{n^3} + 3c_2^2 \cdot \frac{1}{n^2} - 3c_2^2 \cdot \frac{1}{n^3}\end{aligned}$$

Finally,

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{E(S_n^4)}{n^4} &= c_4 \sum_{n=1}^{\infty} \frac{1}{n^3} + 3c_2^2 \sum_{n=1}^{\infty} \frac{1}{n^2} - 3c_2^2 \sum_{n=1}^{\infty} \frac{1}{n^3} \\ &< \infty \quad \text{since these are all } p\text{-series with } p > 1.\end{aligned}$$

Therefore, we've shown

$$P\left(\lim_{n\to\infty} \left| \frac{\sum_{n=1}^{\infty} X_n}{n} \right| = 0\right) = 1 \text{ when } \mu = 0.$$

Now let's suppose $X_1, X_2, X_3, \ldots$ are iid with mean μ. Then set $Y_i = X_i - \mu$. Notice $Y_1, Y_2, Y_3, \ldots$ are iid with mean 0. Consequently,

$$\begin{aligned}1 &= P\left(\lim_{n\to\infty} \frac{\sum_{i=1}^{n} Y_i}{n} = 0\right) \\ &= P\left(\lim_{n\to\infty} \frac{\sum_{i=1}^{n} (X_i - \mu)}{n} = 0\right) \\ &= P\left(\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i - n\mu}{n} = 0\right) \\ &= P\left(\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} = \mu\right),\end{aligned}$$

completing the proof. □

the $\limsup$ **and** $\liminf$ **events.**
Let $A_1, A_2, A_3, \ldots$ be an infinite sequence of events. We define two new events:

$$\limsup_{n\to\infty} A_n = (A_n \ i.o.) = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n,$$

and

$$\liminf_{n\to\infty} A_n = (A_n \ ev.) = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n.$$

Notice that $\omega \in (A_n \ i.o.)$ means that for every $k \geq 1$, $\omega \in \bigcup_{n=k}^{\infty} A_n$, which in turn means for every $k \geq 1$, ω belongs to at least one of the event $A_k, A_{k+1}, A_{k+2}, \ldots$. To summarize, $\omega \in (A_n \ i.o.)$ means ω belongs to infinitely manys of the A_n's. *i.o.* stands for *infinitely often.*

Similarly, $\omega \in (A_n \ ev.)$ means that ω belongs to at least one of $\bigcap_{n=k}^{\infty} A_n$, which in turn means for some $k \geq 1$, ω belongs to *all* the events $A_k, A_{k+1}, A_{k+2}, \ldots$. To summarize, $\omega \in (A_n \ ev.)$ means ω belongs to all the A_n's from some $n = k$ onward, i.e., ω belongs to all the A_n's eventually. *ev.* stands for *eventually.*

Exercise for the student.
Let $A_1, A_2, A_3, \ldots$ be any sequence of events. Show that $(A_n \ ev.) \subseteq (A_n \ i.o.)$.

The Borel-Cantelli Lemma.
Let $A_1, A_2, \ldots$ be a sequence of events.

1. If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \ i.o.) = 0$.
2. If $A_1, A_2, A_3, \ldots$ are independent, and $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(A_n \ i.o.) = 1$.

Proof.

1. Since, for every $k \geq 1$,
$$(A_n \ i.o.) \subseteq \cup_{n=k}^{\infty} A_n,$$
it follows by monotonicity of probability and subadditivity (Boole's inequality)
$$P(A_n \ i.o.) \leq P\left(\cup_{n=k}^{\infty} A_n\right) \leq \sum_{n=k}^{\infty} P(A_n).$$
But, since $\sum_{n=1}^{\infty} P(A_n) < \infty$ it follows
$$\sum_{n=k}^{\infty} P(A_n) \to 0 \text{ as } k \to \infty.$$
Therefore, $P(A_n \ i.o.) = 0$.

2.

$$(A_n \ i.o.)^c = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n^c.$$

By continuity of probability, it follows $P\left((A_n \ i.o.)^c\right) = \lim_{k\to\infty} P\left(\bigcap_{n=k}^{\infty} A_n^c\right)$.

$$\begin{aligned}
P\left(\bigcap_{n=k}^{\infty} A_n^c\right) &\leq P\left(\bigcap_{n=k}^{N} A_n^c\right) \quad \text{for any finite } N \geq k. \\
&= \prod_{n=k}^{N} P(A_n^c) \\
&= \prod_{n=k}^{N} (1 - P(A_n)) \\
&\leq \prod_{n=k}^{N} e^{-P(A_n)} \qquad (*) \\
&= e^{-\sum_{n=k}^{N} P(A_n)}.
\end{aligned}$$

Note that by the MacLaurin expansion,

$$e^{-C} = 1 - C + \frac{C^2}{2!} - \frac{C^3}{3!} + \cdots \geq 1 - C \text{ when } 0 \leq C < 1.$$

So

$$1 - P(A_n) \leq e^{-P(A_n)},$$

which allows us to perform step $(*)$.

Therefore,

$$P\left(\bigcap_{n=k}^{\infty} A_n^c\right) \leq \lim_{N\to\infty} e^{-\sum_{n=k}^{N} P(A_n)} = 0 \text{ since } \sum_{n=k}^{\infty} P(A_n) = \infty.$$

Consequently,

$$P((A_n \ i.o.)^c) = 0, \text{ or equivalently } P(A_n \ i.o.) = 1.$$

□

We immediately have the following

Corollary.
If $A_1, A_2, A_3, \ldots$ is a sequence of independent events, then either $P(A_n \ i.o.) = 0$ or 1 according to whether $\sum_{n=1}^{\infty} P(A_n) < \infty$ or $\sum_{n=1}^{\infty} P(A_n) = \infty$, respectively.

Remark.
What we did earlier in the proof of the SLLN under the assumption of finite 4th moment was to show

$$\sum_{n=1}^{\infty} \frac{E(S_n^4)}{n^4} < \infty.$$

Assume $\mu = 0$. By the Markov inequality for every n, for small but arbitrary $\varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) = P\left(\left|\frac{S_n}{n}\right|^4 \geq \varepsilon^4\right) \leq \frac{E(S_n^4)}{\varepsilon^4 n^4}.$$

Thus,

$$\sum_{n=1}^{\infty} P\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) \leq \sum_{n=1}^{\infty} \frac{E(S_n^4)}{\varepsilon^4 n^4} = \frac{1}{\varepsilon^4} \sum_{n=1}^{\infty} \frac{E(S_n^4)}{n^4} < \infty.$$

Therefore, the Borel-Cantelli Lemma implies, for each integer $j = 1, 2, 3, \ldots,$

$$P\left(\left|\frac{S_n}{n}\right| \geq \frac{1}{j} \ i.o.\right) = 0.$$

The events

$$D_j = \left(\left|\frac{S_n}{n}\right| \geq \frac{1}{j} \ i.o.\right)$$

are increasing as j increases. I.e.

$$D_1 \subseteq D_2 \subseteq D_3 \subseteq \cdots \subseteq \bigcup_{k=1}^{j} D_k \subseteq \ldots$$

and each of these has probability 0. Therefore, by the continuity of probability,

$$P\left(\bigcup_{j=1}^{\infty} D_j\right) = 0.$$

But then this says

$$P\left(\left|\frac{S_n}{n}\right| > 0 \ i.o.\right) = 0.$$

Equivalently,

$$P\left(\left|\frac{S_n}{n}\right| = 0 \ ev.\right) = 1,$$

i.e.,

$$P\left(\lim_{n\to\infty} \left|\frac{S_n}{n}\right| = 0\right) = 1.$$

□

INDEX

D