Math 2030 - Elementary Probability

Class Notes - Student Version

by Aram Dermenjian

July 27, 2021

Week 0

Welcome to Elementary Probability at York University! This semester we'll be going through a ton of various stuff in order to get you better equipped with the tools of probability. These notes are designed to better help you learn the material and guide you through the course.

How to use these notes: These notes are different than most math class notes. They have holes and gaps in them; they are missing information! These are the notes that I will be using in my class lectures (literally) and so it should help you follow along. The purpose of these notes is so that you don't have to write everything down; you only need to fill in the gaps giving you more time to listen to lectures and better understand the material. If you need to add extra notes, there is *tons* of room in the margins, use them! These are YOUR notes. I've also added links to Wikipedia in case you want another reference to the terms and definitions we go over. If you feel using these notes would hinder your learning experience, then *please* use another system. The point is to learn; so use whatever works best for you.

Before starting a lecture I'd recommend quickly glancing over the notes (5-10 minutes) in order to get a brief idea of what we'll be covering in class. These notes are a supplement to the "official textbook". They will cover roughly 400 pages of the "official textbook" over the next 12 weeks. That means, roughly 33 pages per week! Remember to review these notes periodically so that you don't forget terms and examples.

Good luck, I believe in you!

-Aram

Week 1

13 May 2021

We start this class at the very beginning: what is probability? When we think about probability, we normally think that it's the study of chance. Such as, what are the chances that I'll win the lotto or that our favourite team will win the tournament. But we can also use probability to figure out whether the information we are looking at is accurate, how different events can interact with one another, and so much more.

Since, in mathematics, we like to be precise when we talk about the world, we will be working towards taking the word "chance" and making it as precise as possible.

1 Events

Normally when we think of chance we think of flipping a coin, rolling a dice or playing the lotto. All of these are examples of chance which come from equally likely outcomes. For example, if we look at flipping a coin, I have $\frac{1}{2}$ chance that it will turn out to be heads and $\frac{1}{2}$ that it will be tails. An equal $50\% - 50\%$ for both outcomes. We start our exploration of probability with examples such as the above; examples where the outcome is equally probable.

So say I want to roll a dice and I want to figure out my chances of rolling a number. We inherently know the chance, but how do we do this mathematically? **First,** we need to take all possible outcomes and put them into a set. This set is called the *sample space*, and we (usually) denote it with an Ω. In our example of rolling a six-sided die, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The elements in our sample space Ω are normally called *outcomes*. Something that might or might not occur based on the outcome are called *events*. Using the dice, we might have events such as:

- The dice rolls a six.
- The dice rolls an even number.

Wikipedia: Sample Space
Note that in the book by Pitman, "sample space" is called an *outcome space*. The two are the same thing.
Wikipedia: Outcome

- The dice rolls a number less than 3.
- The dice doesn't roll a 4.

Our aim is to study what the probability or chance of certain events occurring is.

Example 1.1 Let's do an example. Say we're about to do a raffle in class and there are 200 students. We draw one name from a hat for who will get an $A+$ in the class, no questions asked. What is the sample space of the raffle? $\Omega =$ _____.

Try and list some events for this space:

-
-
-

So what is an "event" mathematically? An *event* is a subset of our sample space Ω. For example, "the dice rolls a six" is represented by the subset $\{6\}$. The event "the dice rolls an even number" is represented by the subset $\{2,4,6\}$. How is the event "the dice rolls a number less than 3" represented? _____.

Wikipedia: Event

Sometimes, language actually helps us create these subsets!

"A or B" means $A \cup B$.

Or: Say I take the event: "The dice rolls a six or a two". This is really the combination of two potential events: "the dice rolls a six" and the event "the dice rolls a two". Since both are possible, we can represent this as: $\{6\} \cup \{2\} = \{2,6\}$. In other words, for most purposes, the word "or" can be thought of as a union of two sets.

"A and B" means $A \cap B$. Note that the book uses AB to mean $A \cap B$.

And: What about "and"? Let's test it out with our event above my switching the word "or" with "and" - "The dice rolls a six *and* a two". Inherently, if we're rolling a die one time, this is impossible. You can't have two numbers show up by rolling a die so we expect the sample space to be empty. In terms of sets, we can think of "and" by intersecting sets: $\{6\} \cap \{2\} = \emptyset$. When we see that an event is empty (aka equal to $\emptyset$) then we say that the event is *impossible*. The opposite of an impossible event is a *certain event* and is when the event is Ω.

"not A" means $\Omega \backslash A$.
This is sometimes known as the "complementary event".
Wikipedia: Complementary event

Not: The final word we'll quickly go over is the word "not"; in other words, what is the chance of something *not* happening. An example of this is the event "the dice doesn't roll a 4". What we really want here is to take our set to be $\{4\}$ and remove it from our sample space in order to be left with the actual possible outcomes. In other words, we take the complement of our sample space: $\Omega \backslash \{4\} = \{1,2,3,5,6\}$.

Example 1.2 Let's try the above with the stereotypical example of flipping a coin. The sample space here is given by $\Omega =$ _____. What subset of Ω is represented by the following events:

- The outcome is heads _____
- The outcome is heads or tails ___
- The outcome is heads and tails __
- The outcome is not head _____

Here is a table of some useful conversions from English to Set theory/Probability.

Human language	Notation	Set language
Sample space	Ω	Collection of outcomes
Event that some outcome in A occurs	A	Subset of Ω
Not A	$A^c = \Omega \backslash A$	Complement of A
A and B	$A \cap B$	Intersection
A or B	$A \cup B$	Union
A, but not B	$A \backslash B$	Difference
Either A or B, but not both	$A \triangle B$	Symmetric difference
If A then B	$A \subseteq B$	Inclusion
Impossible event	$\emptyset$	Empty set
Certain event	Ω	Whole sample space.

2 Probability

Now that we've discussed events, we want to know what the probability of a certain event occurring is. If all outcomes in a finite set Ω are equally likely then the probability of an event $A \subseteq \Omega$ to occur is:

Wikipedia: Probability

Example 2.1 Let's look at the four events we had at the very beginning with rolling a dice. We know that $\Omega = \{1, 2, 3, 4, 5, 6\}$ and so $|\Omega| = 6$.

- "The dice rolls a six." – First we need to calculate our event A. In this case $A = \{6\}$. So we have

$$P(A) = \frac{\quad}{\quad}$$

- "The dice rolls an even number." – This statement is equivalent to "The dice rolls a 2 or a 4 or a 6". Recalling that "or" means union, we have $A = \{2\} \cup \{4\} \cup \{6\} = \{2, 4, 6\}$. Therefore

$$P(A) = \_\_\_$$

- "The dice rolls a number less than 3." –

- "The dice doesn't roll a 4." – We saw before that

$$A = \{1, 2, 3, 5, 6\} = \{4\}^c = \Omega \backslash \{4\}.$$

Therefore

$$P(A) = \_\_\_$$

Example 2.2 What happens if we flip multiple coins instead of one? Let's say we have 3 coins and we flip them one at a time. What is our sample space?

Each coin can be heads or tails, and since we are flipping them one at a time we can put them in the order of when we flipped them. In other words we can think of each outcome as a list of three outcomes:

Notice how each outcome has three values and they are ordered based on which coin it is referring to. In other words (H, H, T) is not the same as (T, H, H) even though they both represent that you got two heads and one tails. *The* _____ *matters in this case since we are flipping them one at a time.* Therefore our sample space Ω contains the 8 elements above (*i.e.*, $|\Omega| = 8$).

Let's look at some events. Say our event is "Flip heads twice." First, we must notice that this statement is ambiguous! Does the statement mean we get heads exactly two times or does it mean we flip heads at least two times? (So three heads would be ok?) *Every word is important in probability.* Let's try both and see the difference.

Flip heads exactly twice: In this case we have

$A =$

which are all outcomes in our sample space Ω where heads appears two times. Therefore our probability is given by:

$$P(A) = \frac{\quad}{\quad}$$

Flip heads at least twice: In this case we are really asking "Flips heads twice or flips heads three times". So our event is given by

$A =$

which are all possible outcomes. Therefore our probability is given by:

$$P(A) = \frac{\quad}{\quad}$$

Another event we might want to look at is "We flip heads first". If you think about it for a second, you'll realize we're just asking what the first flip's outcome is, but we'll be pedantic and do things slowly. Our set A should contain every outcome whose first entry is H:

$A =$

Therefore our probability is given by:

$$P(A) = \frac{\quad}{\quad}$$

Example 2.3 The following example is a common student mistake! Let's see what happens when you think about these flips differently. If we have 3 coins and we flip them all at once. We might be tempted to say that we can just look at the sets themselves in order to determine the probabilities. In this case we would set our sample space to be:

$\Omega =$

In other words, we only have 4 potential outcomes instead of 8. What happens when we look at the events from the previous example?

"Flip heads at least twice": Our new event is given by

$A =$

This gives us a probability $P(A) = \frac{\quad}{\quad}$ which is the same as before! We might be tempted to say that order doesn't matter, until we look at one of the other events.

"Flip heads exactly twice": Our new event is given by

$A =$

giving us a probability $P(A) = \_\_$. This is definitely not the same as before!

So what are we doing wrong? First, we need to realize that flipping coins are all independent of one another. In other words, the result on one coin does not influence the result on another coin. This means they are "independent" which is something we will cover in more depth later on. Because our coins are independent, looking at coin tosses as sets is problematic because it doesn't take into account variations within the set. In other words, the set $\{H, H, T\}$ doesn't take into account that the tail flip could have come from the first coin or the second coin or the third coin! Notice how I said "the first coin or the second coin or the third coin". What this set actually represents is a *union* of our three lists from the previous example, but we randomly decided to combine them into a set where order doesn't matter.

This is probably the hardest part about probability: making sure that your sample space and your events are accurately describing the problem. It is very easy to merge things together because you didn't think about all possible variations beforehand.

3 Counting

We're going to take a small detour and talk about counting since we are noticing that set theory and probability and intimately related. Most of these seem trivial, but it's good to be mathematically precise when we talk about them.

We let $|A|$ denote the number of elements in a set A. Recall that a $1-1$ *correspondence* is

The correspondence rule: If we have two sets A and B and we can make a $1-1$ correspondence between the elements then

We say that two sets A and B are *disjoint* if $A \cap B = \emptyset$. If we have multiple sets $A_1, A_2, \ldots, A_n$ then we say they are *mutually disjoint* if every pair of sets are disjoint, *i.e.*, $A_i \cap A_j = \emptyset$ for all i, j.

The addition rule: If we have a set A that can be split into a collection of mutually disjoint sets $A_1, A_2, \ldots, A_n$ then

The multiplication rule: Suppose we have k choices to be made one after another with exactly n_j options at each choice with $j \leq k$. Then the total number of options which can be made are

Sequences: A *sequence of length k of elements of S* is an ordered list which has k elements each coming from the set S. Each component is chosen without consideration of the other ones and so there are $|S| = n$ options to fill in each component giving us ____ total sequences possible.

Example 3.1 For example if $S = \{A, B, C\}$ and we look at $k = 4$ sequences we'll have $3^4 = 81$ different sequences. Here are some:

Orderings: An *ordering of k out of n elements* is a sequence of k elements where every entry is different. This means that each time we chose an element, it can no longer be selected! So we have n options for the first entry, $n-1$ options for the second entry, etc. This means there are

total options. Recall that$n$ factorial is given by $n! = n(n-1)(n-2)\dots 2 \cdot 1$. Then we can represent the above as _____. $0! = 1$

Example 3.2 For example if $S = \{A, B, X, Y\}$ and we let $k = 2$ then we have the following $\frac{4!}{2!} = 4 \cdot 3 = 12$ options

Notice how the order matters!

Permutation: A *permutation of an ordering of k out of n elements* is a reshuffling of the entries of an ordering.

Example 3.3 For example if I have the ordering (A, B, C) then we have six permutations:

Notice how each permutation has the same content, but the order in the orderings is different.

In this case we have $3 \cdot 2 \cdot 1$ total options. Normally when we see something like this, we use the factorial symbol. In this case $3! = 3 \cdot 2 \cdot 1$.

For an ordering with k elements we have ___ permutations.

Combinations: Say that we don't actually care about order. So (A, B) and (B, A) are the same thing. In this case we can look at our ordering and divide by the number of permutations each ordering has. A *combination of k elements out of an n element set S* is a subset of S with k elements. We end up with $\frac{n!}{(n-k)!k!}$ which we normally denote by _____. This is called the *choice function* or the *binomial coefficient*. You can read it as "n choose k" because we are taking a set of n elements and we are choosing k of them.

Wikipedia: Binomial coefficient
The term "binomial coefficient" is coming from algebra where $(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$.

Example 3.4 For example if $S = \{tree, fern, sprig, grass\}$ then we have the combinations of 3 elements given by:

$$\{tree, fern, sprig\}, \{tree, fern, grass\}, \{tree, sprig, grass\}, \{fern, sprig, grass\}$$

Notice that all we did was we *chose three out of four* elements: $\binom{4}{3} = \frac{4!}{3!1!} = 4$.

Subsets: Finally, suppose I have an n element set S and I want to look at all possible subsets of S. Then for each element I need to decide whether that element is in our subset or not. In other words, I have 2 options for each element. This gives ____ different possibilities.

Example 3.5 For example if $S = \{A, B\}$ then my subset options are:

$$\emptyset, \{A\}, \{B\}, S$$

giving us $2^2 = 4$ total subsets.

We normally denote by 2^S the _____. In other words:

4 Odds

Another way to discuss an event is to describe the odds of an event happening which is often used in gambling. (Think of when people say "the odds are 3 to 2 that the person will win".)

Odds are normally just another representation of probability that compares the chance of something occurring vs the chance of something not occurring In mathematical terms, the *odds in favour* of something happening are $|A|$ to $|A^c|$. Alternatively, the *odds against* something happening are $|A^c|$ to $|A|$. Usually these are denoted as $|A| : |A^c|$

Wikipedia: Odds

Example 4.1 For example, if we roll a six sided dice, and we want to find the odds of rolling a 4.

We use odds super rarely in this class as it's just a different representation of probability itself. It is something that you should know though as this is an extremely common way of representing probability in the real world.

Example 4.2 Say you're watching a horse race and they say horse A has odds $7 : 1$ of winning. What is the probability that the horse will win?

5 Interpreting probability

So we talked about the (extremely super) basics of probability, but a question we ask next is how should we interpret a probability? In other words, when we know the probability of something occurring, how do we want to use/interpret that probability? There are two main ways of interpreting probability called "frequency" and "subjective" interpretations. There are other interpretations such as "propensity", "logical" and "predictive", but those you can look up on Wikipedia or another book.

Wikipedia: Interpretations

Frequency interpretations ask _____. When we look at probability from this standpoint, we call it *frequency probability*.

Wikipedia: Frequency probability

Usually we use this interpretation when something happens more than once. If you find yourself asking "how often will x occur", then you're thinking in frequency terms.

In real life, we normally test probability by running examples to see how often things actually happen in order to make sure we didn't forget something. For this, we make a prediction using probability and then we run simulations to record the relative frequency. The *relative frequency* is a proportion measuring how often/frequently something occurs in a sequence of observations. We normally record these observations over time and see what value they trend to.

Example 5.1 Say we want to verify that the chance of rolling a 5 on a six-sided die is $\frac{1}{6}$. What we would do is roll the die once and record if it was a 5 or not. We then roll it again and record if it was a 5 or not. We repeat this process for an extremely long amount of time, always recording whether 5 occurred or not. This gives us a list of how often 5 occurred up to a point. For example say we were super lucky and saw 5 appear the 3rd, 4th and 7th time we rolled a dice out of ten times. Our list would look like:

Notice that the numerator increased by one each time I had a success (*i.e.*, I rolled a 5) and the denominator increased for each time an attempt was made.

After our simulations are done, we can plot these results on a graph to see what they look like. Our simulation gives us points $(i, \frac{x_i}{i})$ where i is the ith roll and x_i is the number of cases up to that point. Using our data from above we would have the points:

I used a computer to simulate rolling a die 500 times and the resulting plot looks like this:

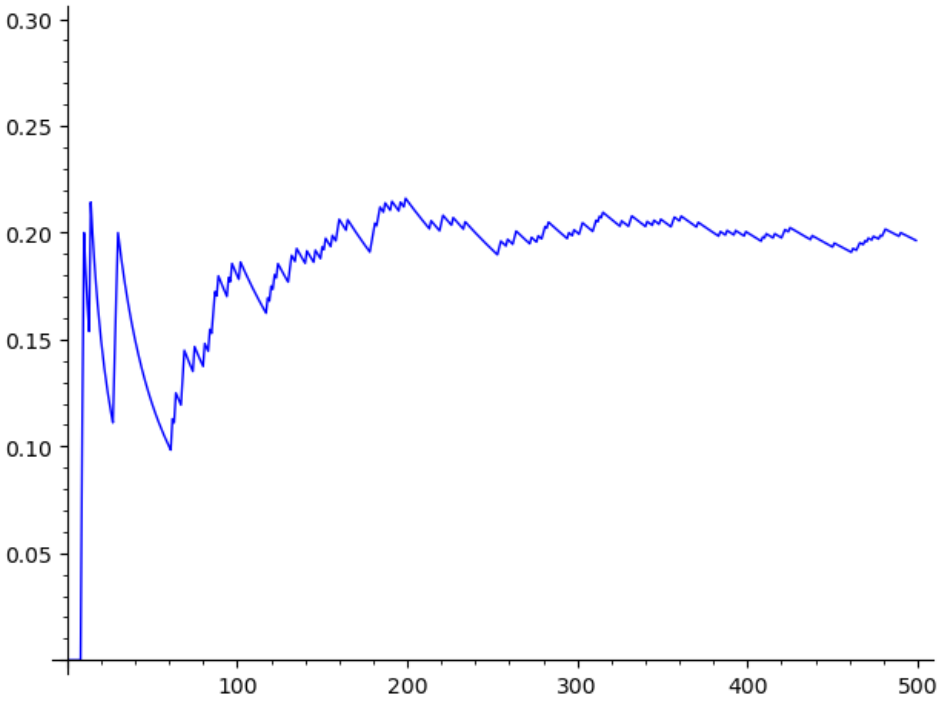


Figure 1.1: Frequency of 5 on a six-sided dice out of 100 trials

Notice how we never get exactly $\frac{1}{6} \approx 16.6\%$, but we get closer and closer as time goes on.

Subjective interpretations ask us to calculate chance using _____. For example, during an election we might want to know whether our preferred candidate has a chance of winning or not or which horse is going to win a horse race or what the chances of surviving a medical procedure are. When looking at probability from this standpoint, we refer to it as *Bayesian probability*.

Wikipedia: Bayesian probability
The textbook refers to Bayesian probability as subjective probability.

Since these types of interpretations are normally based on opinions and there are a lot of uncertainties involved, we won't go into this type of interpretation deeply. For now, it suffices to know that it exists and what it means.

6 Distributions

We're now going to take probability and put it neck deep in mathematics. So far what we've been doing is starting with some set Ω of possible outcomes which we called the sample space. We then took events and we represented them by subsets A of Ω. Finally, we created a probability function P which took A and assigned it a number.

We'll now formalize this and try and make it more mathematically precise. We let Ω be some set of outcomes. Note that we don't assume Ω to be finite nor do we assume that every outcome is equally likely to occur. Let 2^Ω denote the set of all subsets of Ω and let $A \in 2^\Omega$, *i.e.*, A is a subset of Ω. We define the probability map to be the map $P : 2^\Omega \to \mathbb{R}$ which sends every subset to some real number.

Although all of this makes mathematical sense in set theory, we want to ask, how do we know if this represents probability or not? For this, we need to define some set of rules which ensure that P behaves "nicely". One easy example is that for every A we want $P(A)$ to be positive! (What does negative probability even mean?) What other rules do we need?

Let's consider multiple events happening. Since we're thinking of events as sets, let's break down what happens when sets interact with one another. We've already seen union, intersection, etc. but now we'll introduce something called a partitioning. We say that two sets A and B are *disjoint* if $A \cap B = \emptyset$. We use the notation $A \sqcup B$ to mean $A \cup B$ where A and B are disjoint. An event A is *partitioned* into n smaller events if:

$$A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_n.$$

For us, event and subset are synonymous. Also, remember that $\cup$ just means "or".
We can also technically partition A into an infinite number of subsets! We'll look at this in later classes.

Since we are breaking apart A into smaller chunks, one things we want is that the sum of the little chunks should equal the probability of A itself. This gives us a second rule:

$$P(A) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

Another rule we might want is that if $A = \Omega$ then our probability should be 100%.

In particular, we end up having three rules, which together we call the *rules of proportion and probability*.

- **Non-negative**: $P(A) \geq 0$
- **Addition**: If $A_1, A_2, \ldots, A_n$ is a partition of A then

$$P(A) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

- **Total one**: $P(\Omega) = 1$.

A *distribution over* Ω is a function of subsets of Ω satisfying these three rules.

Wikipedia: Distribution

With this we can be a little more stringent with our language translations from before.

Theorem 6.1 (Complement Rule) *The probably of the complement of an event A is*

$$P(\text{not } A) =$$

Proof.

□

Note that the book assumes $A \subseteq B$, but we don't make that assumption.

Theorem 6.2 (Difference Rule) *The probability that B will occur and A will not is given by:*

$$P(B \text{ and not } A) =$$

Proof.

□

Theorem 6.3 (Inclusion-Exclusion) *Let A and B be two subsets of Ω. Then*

$$P(A \cup B) =$$

Proof.

□

Example 6.4 Say I have a bag of marbles and I tell you that 15% of the marbles are blue, 12% are red and 5% are both red and blue. If I pull out a marble, what are the chances that the marble is not blue?

If I pull out a marble, what are the chances that the marble is red, but not blue?

If I pull out a marble, what are the chances that the marble is either

blue or red?

Example 6.5 Say I roll a 6-sided die twice. What is the distribution of the sum of the values?

Recall that the sample space for rolling a die twice can be represented by ordered pairs (i, j) and that we end up with 36 possible outcomes. To compute the distribution of the sums, we need to figure out how many times each sum appears.

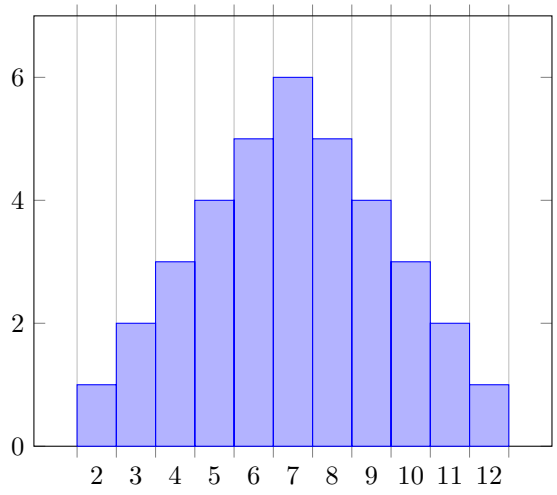
We know that to get a 2 we need to roll a 1 followed by a 1 (represented by $(1, 1)$). Since this is the only way we know $P(2) = \frac{1}{36}$.

Similarly, to get a 7 we have the following 6 options

$$(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)$$

Therefore, we get the following distribution:

$$P(2) = \frac{1}{36},\ P(3) = \frac{2}{36},\ P(4) = \frac{3}{36},\ P(5) = \frac{4}{36},$$
$$P(6) = \frac{5}{36},\ P(7) = \frac{6}{36},\ P(8) = \frac{5}{36},\ P(9) = \frac{4}{36}$$
$$P(10) = \frac{3}{36},\ P(11) = \frac{2}{36},\ P(12) = \frac{1}{36}$$



6.1 Bernoulli p distribution

Wikipedia: Bernoulli p distribution

Probably the easiest distribution to understand is the *Bernoulli distribution*. It is basically a generalization of _____. We suppose that Ω has exactly two outcomes $\{A, B\}$. We let $P(A) = p \in [0, 1]$ and we let $P(B) = 1-p$, which also is contained in $[0, 1]$. Notice that $A^c = B$ and $B^c = A$.

Let's verify that this actually gives a distribution. First, notice that every subset gives a probability between 0 and 1. Next, notice that the only set that can be partitioned is Ω with the partition A and B. Also notice that

$$P(\Omega) = P(A) + P(B) = p + 1 - p = 1.$$

Since all three points of the rules proportion and probability are satisfied, this is a distribution!

6.2 Discrete uniform distribution

Wikipedia: Discrete uniform distribution

Note that the book uses "uniform distribution on a finite set" to mean discrete uniform distribution.

If $n = 2$, then the uniform distribution is the same as the Bernoulli $\frac{1}{2}$ distribution.

A *discrete uniform distribution* is a generalization of _____. In essence we let $\Omega = \{1, 2, \ldots, n\}$ be a set of n elements. We suppose that the probability of choosing any one of them is given by $\frac{1}{n}$, *i.e.*, each outcome is equally likely.

To verify this is a distribution, we first notice that $P(A) = \frac{|A|}{n}$ for every $A \subseteq \Omega$ and that $P(A) \in [0, 1]$ always. Partitioning can be proved inductively, and the final rule can be deduced from above.

6.3 Continuous uniform distribution

Wikipedia: Continuous uniform distribution

Note that the book uses "uniform (a, b) distribution" to mean continuous uniform distribution.

A *continuous uniform distribution* is a generalization of the _____ to the set of real numbers. Here the idea is that we take an interval (a, b) (where $a < b$) and we pick a point randomly in the interval. (We assume that every point has an equal chance of getting picked). In this case we can't ask questions like "What is the probability that we pick the point x" because the chance is 0. (The reasoning for this is discussed in future lectures). Instead we ask questions like "What is the probability that the point is between x and y where $a < x < y < b$". The probability in this case is given by $\frac{y-x}{b-a}$. If $a = 0$ and $b = 1$ then we call this the *standard uniform distribution*.

Week 2

20 May 2021

7 Conditional probability

A lot of times we want to know the probability of something occurring while also having some information about what has already happened. As a quick example, imagine your friend has just rolled two dice, but they kept the results from you. They ask you what are the chances that they got snake eyes (two 1s). They then let you see one of the dice and it's a 1! We now have more information and the question becomes whether this new information will change the probability that they rolled two 1s. This is the idea of conditional probability.

Example 7.1 Let's actually go into this example a little deeper and see what happens. First we start by figuring out the initial probability that your friend rolled two 1s on their die. Since each dice has 6 sides there are 36 different possibilities:

$$\begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array}$$

The set of these thirty-six outcomes gives us the sample space Ω. Our friend is asking what the chances are that they rolled two 1s. Looking at our sample space, we notice that $A = \{(1,1)\}$ which means our probability is _____.

At this point your friend shows you that one of the two die is in fact a 1! What is the probability that the other dice is a 1 as well? We might think that since the other dice has 6 options, that the chance should be

Common student mistake

$\frac{1}{6}$, but that's not exactly correct. Your friend has a *choice* on which die they showed you. It could have been the first one or the second one and so it's not as simple as $\frac{1}{6}$. Instead, what we have to do is look at all of our original 36 possibilities and ask which ones have a 1 in them to see our new sample space. We're left with:

So our sample space has 11 elements and only 1 of them is $(1,1)$. Our probability is $\frac{1}{11}$ which is much less than $\frac{1}{6}$.

How do we represent this mathematically? Notice that our event A never changed. The event stayed as "two 1s" and so $A = \{(1,1)\}$ throughout the whole process. Instead, we added more information. Our friend told us that "one of the dice is in fact a 1". Therefore we have a new event B which is "one of the dice is 1" which gives us the set of 11 elements from above. Our sample space is therefore B and our new set is whatever is in both A and B.

In other words, the *conditional probability of A given B* is

Wikipedia: Conditional probability

Although this is nice, we want to try and work with probabilities as much as we can. So what we do is we convert the previous formula into probabilities:

Another way to write this is the following:

We call this formula the *multiplication rule*. Intuitively what this is saying is if the event B happens around $1/2$ the time and if $1/4$ the time that B happens the event A happens , then A and B happen about $1/8$ the time.

7.1 Tree diagrams

We can actually represent all of this with what are known as tree diagrams. A *tree diagram* is basically a scheme which allows us to graphically represent sample spaces for conditional probability. The best way to see a tree diagram is through an example.

Example 7.2 Suppose we have three bags of chocolate, each one containing three different types of chocolate: White (W), Dark (D) and Milk (M). But we know that the the quantities in each bag are different! If the first bag has 2 whites and 1 milk, the second bag has 2 dark and 2 milk and the third bag has 1 white, 1 dark and 2 milk, what is the probability that we will get a dark chocolate if we randomly draw a chocolate from a random bag?


Common Student Mistake: Before we use conditional probability, let's talk about a common student mistake at this point. A common mistake is to just add up all of the quantities ($2+1$ white, $1+2+2$ milk and $2+1$ dark) and say that since there are 3 dark chocolates and 11 chocolates in total, you have a 3/11 chance of getting a dark chocolate! This is wrong because you are completely forgetting about the bags! If you chose the first bag, you have absolutely no chance of getting dark chocolate. It's important to keep track of *all* the information given in a problem. Every word counts.

With the warning out of the way this is a perfect time to use conditional probability and tree diagrams. For this we set up our diagram in the following way:

Example 7.3 Let's look at another example that might deal more with real life. Say you're midway through a class and want to know what your

chances are of passing the class. You need to pass both exams to pass the class. You've studied hard for the exam coming up and you're sure that you have a 95% chance of passing the first exam. The second exam is harder to predict though and you think that if you pass the first exam, then you have a 90% chance of passing the second exam, but if you fail the first exam your chances will go down to 75%. What is the chance you will pass the class?

We start off by making a tree diagram as before.

What are your chances of passing at least one exam? What about passing exactly one exam?

This second example is a good time to talk about average conditional probabilities. We had set our event A to be "pass exactly one exam" and we can condition that on the event B which is "pass the first exam".

What does this look like mathematically? We actually already saw this above! We had

But notice that this is just a partition! We can extend this argument to any partition to get the *law of total probability*.

Wikipedia: Law of total probability

Theorem 7.4 (Law of total probability) *If $A_1, A_2, \dots A_n$ is a partition of Ω then*

Note that the book calls this the *Rule of average conditional probabilities*.

8 Independence

In the examples we saw in the previous section, our second event was always *dependent* on our first event. But this is not always the case. Think of rolling two die. If I roll the first dice, it doesn't magically change the probability for the second dice. Since the second die is not dependent on the first we say the two events are independent.

Mathematically, we think of this as the probability of an event A not changing no matter whether B occurs or not:

In this case say A and B are *independent*. A much more simple way to view this definition is the following:

Wikipedia: Independence

Working this around we have:

This gives us the *multiplication rule for independent events*

The multiplication rule is often taken to be the definition of independence (for example, Wikipedia uses this as the definition of independence). Any of the three definitions I have given are all valid responses for the definition of independence.

Example 8.1 Let's look at flipping a coin twice as an example. We first will draw the tree diagram:

We easily see that our events are independent just by looking at the diagram.

9 Sequence of events

We started looking at doing one event followed by another, but what happens if we have a chain of multiple events? Say I flip 10 different coins, or I draw 7 cards, or I pull out 34 bars of chocolate from a bag. How do we find the probability of a sequence of events?

Let's first start off slow. If we have two events we saw:

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

How about if we have three events A, B, and C? Since we're going in order C is dependent on A and B, *i.e.*, $A \cap B$. So we have:

$$P((A \cap B) \cap C) = P(A \cap B)P(C \mid (A \cap B)) = P(A)P(B \mid A)P(C \mid (A \cap B))$$

We should start seeing a pattern at this point. It turns out, we can keep doing this forever and we get a *multiplication rule for n events*

Example 9.1 Let's look at a quick example of how this might work using our tree diagrams. Suppose we want to flip a coin two times with the condition that if we flip a tails, then we draw a number between 1 and 3 out of a bag. Once we've drawn a number out, then we stop. If the second flip is a heads, then we consider that as drawing a 0. The tree diagram would look like:

What is the probability that we flipped exactly 1 tails?

What is the probability that we pulled a 1 out of the bag?

We can also extend the notion of independence to an arbitrary sequence of events. If we have n events $A_1, A_2, \ldots, A_n$, then we say they are *independent* if they are pairwise independent. What this means is for any two events A_i and A_j then A_i and A_j are independent. In this case, the multiplication rule gives us:

An easy example of this is to just think of a coin toss. If you flip multiple coins the coin flip of one coin doesn't influence another coin's result. So the order that you flip the coins doesn't matter.

Example 9.2 We finish with an extremely standard example of how things in probability have unexpected results, can be complicated and benefit from having structure. This problem is called the *birthday problem*. It's a simple question: Say there are n people in a room. What is the probability that at least two people have the same birthday? (We ignore leap days... sorry)

10 Bayes' Rule

Let's go back to the chocolate in a bag example from before. Recall that we had three different bags which contained different chocolates. The first bag has 2 whites and 1 milk, the second bag has 2 dark and 2 milk and the third bag has 1 white, 1 dark and 2 milk. We were told we first randomly chose a bag and then pick out a chocolate. The question we asked at the time was, what are the chances the chocolate is dark.

Now, we're going to ask a slightly reverse question. Say that I randomly pick a bag and pull out a chocolate and I show you that I pulled out a dark chocolate. I then ask you, which bag do you think I pulled the chocolate out of. Basically, we want to try and go backwards with probability. If you look at the bags, the second bag has the most dark chocolates so we'd predict that the second bag was the bag I most likely pulled the chocolate from.

Let's analyze this mathematically.

Let's try and generalize this. Let A be our event "pulled a dark chocolate". Then we have three "prior" events which are the three bag choices: B_1, B_2, B_3. What we wanted to calculate was _____. To do this we did:

This is called *Bayes' Rule*.

Wikipedia: Bayes' rule
Bayes' rule is sometimes called Bayes' theorem, Bayes' law or Bayes-Price theorem.

Note that we're starting to notice that $P(A \mid B)$ and $P(B \mid A)$ are both things we can measure! So it makes sense to start differentiating which event comes before another event if we have a sequence of events. If event A happens before event B then $P(A \mid B)$ is called the *posterior*

Wikipedia: Posterior probability
Wikipedia: Likelihood
Wikipedia: Prior probability

probability of A given B and $P(B \mid A)$ is known as the *likelihood of B given a fixed A*. The probability of event A (*i.e.*, $P(A)$) is referred to as the *prior probability*.

Recall that earlier I asked the following question: "Say that I randomly pick a bag and pull out a chocolate and I show you that I pulled out a dark chocolate. What are your chances of guessing which bag I pulled the chocolate out of?" Let's next say that I'm being manipulative and I pose the following question instead of the one from before. "Say that I pick a bag and pull out a chocolate and I show you that I pulled out a dark chocolate. What are your chances of guessing which bag I pulled the chocolate out of?"


Common student error: It's the same as before!

It's actually not the same as before! What is different about the second question? _____.
Usually when we use the word "randomly" we are making the *implicit assumption* that all outcomes are equally likely to occur. So when I "randomly chose" a bag, each bag has $\frac{1}{3}$ chance of being chosen. But what if it's no longer random? This actually makes the problem *much* harder because you no longer know the values of $P(B_i)$ from above. At this point, it's better to keep them as variables since you don't have enough information to solve the problem. This gives the following:

Week 3

27 May 2021

11 The binomial distribution

We're now going to look at a generalization of flipping a coin (aka, doing an action multiple times where the actions are independent of one another). We saw this partially when we talked about the Bernoulli p distribution.

Suppose that I have a fair coin and I flip it multiple times. I don't necessarily care about the order that the flips come in, but I do care about the end result. For example, if I flip the coin four times and I get (H, T, T, H), I just care that I got two heads and two tails. Let's see how many possibilities we have when keep flipping the coin.

The first time I flip it I can either get a heads or a tails:

The second time I flip, I again can either get a heads or tails. But since we don't care about order (H, T) and (T, H) are the same. So we have

Let's flip a third time. We end up getting:

Notice that this just gives us the binomial distribution which we talked about in the counting section in the first week. We can draw these numbers using *Pascal's triangle*:

Wikipedia: Pascal's Triangle

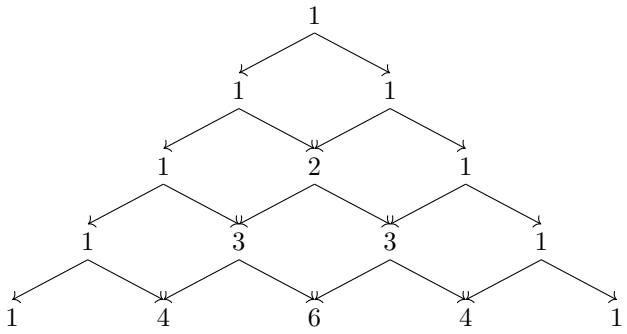


Figure 3.1: Pascal's Triangle

We can now use this to figure out probabilities given an arbitrary (not necessarily fair) coin. If the coin has probability p of landing on heads, then we can label each edge with p for the probability we flip a heads and a $q = 1 - p$ to represent a tails flip. So for example, if I flip 3 times, the probability that I will get 2 heads and 1 tails is $3p^2q$.

The name "binomial" comes from the expansion of the binomial $(x + y)^n$. For example:

We can keep extending Pascal's triangle infinitely in order to calculate any n that we want. And so we might wonder, how does this help with flipping coins? This is where it helps to recall that the numbers in Pascal's triangle are given by the choice function. In particular if we go to the nth row and look at the kth number, that number is equal to $\binom{n}{k} = \frac{n!}{(n-k)!k!}$. In other words, if I flip my coin n times and k of them turn out to be heads then we know the probability is equal to:

This is known as the binomial distribution.

To be more precise, if we have n independent trials with probability of p that a particular trial is successful, then the *binomial probability formula* is given by:

For a fixed n and p, the binomial probabilities for all k give a probability distribution over the numbers $\{0, 1, \ldots, n\}$ called the *binomial distribution*.

Wikipedia: Binomial Distribution
Note that the book uses the term "Binomial (n, p) distribution".

Notice how we have 0 in our distribution! This is why we let $0! = 1$; or else we couldn't work with factorial properly.

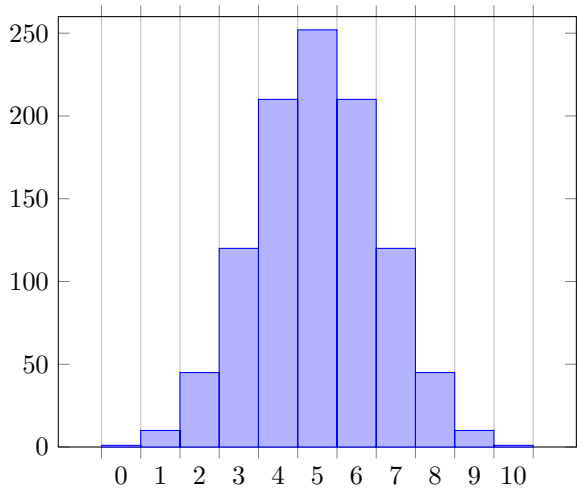
Example 11.1 Let's look at an example of something other than flipping a coin. Say we have a six sided dice and we want to figure out what the probability of getting 3 sixes out of 7 rolls. First, we know that each roll is independent and so we know we can use the binomial probability function. We know that $p =$ ___ and since we are doing 7 rolls we have $n =$ __. As we want 3 sixes we also know that $k =$ __. Therefore we have:

Example 11.2 Now we get a little more complicated. Say we're in a coin flipping tournament and there are five contestants. Each contestant must flip six coins and are allowed into the next round of the competition if they get more than three heads. What are the chances that at least three contestants proceed to the next round?

12 Consecutive Odds Ratios

It turns out that for the binomial distribution, odds are actually a nicer way to look at things. Remember that odds compare two things and say what the ratio between two things occurring is. (So "Horse A has $2:3$ odds in favour of winning against horse B" means out of ever 5 races, horse A will win 2 and horse B will win 3 on average) We're going to look at consecutive odds in the binomial distribution and see what they give.

Example 12.1 Let's look at the distribution when $p = \frac{1}{2}$ and $n = 10$. We end up with the following:



The distribution is given by:

Let's compare them consecutively:

It's easy to see that this generalizes generally:

The above only gives us consecutive odds if the probability of success is equal to $\frac{1}{2}$, but it's not hard to see how this expands to arbitrary probability.

Theorem 12.2 (Consecutive odds for the binomial distribution) *For independent events with a probability of success p. Then the odds of k successes relative to $k-1$ success are given by:*

Example 12.3 Let's see how this helps. Say that I have hat with five numbers in it and every week I pull out a number for seven weeks. What is the probability distribution of pulling out the number 4.

13 Most probable outcome

Notice how in the previous example $P(1)$ had the highest value. This implies that the number 4 has the highest probability of appearing exactly 1 time. So it makes sense to ask if we can generalize this. If I have n independent trials and each trial has probability p of being successful then we might expect to have roughly ____ successes. (For example, if I flip a coin 4 times then I expect _____ heads.) If we let m be the trial which has the highest probability, *i.e.*, $P(m) > P(k)$ for all $k \neq m$, we call m the *mode* of the probability. We end up with the following:

Theorem 13.1 (Mode of the binomial distribution) *If $p \in (0,1)$ where p is the probability of success of an independent trial and we suppose that we perform n trials. Then the mode of the binomial distribution is given by m where*

Wikipedia: Expected value

A related idea in probability is the expected value. The *expected value* is the expected number of successes. . This is different than the mode which is the most likely number of successes. The expected value is defined to be np and it is usually denoted by μ for the binomial distribution.

The book uses the term *mean* as the expected value of a binomial distribution.

The difference between the two is best seen through examples.

Example 13.2 First let's see when they are the same. If $n = 4$ and $p = \frac{1}{2}$ (aka we flip a coin 4 times) we expect the number of successes to be:

This is the expected value.

The most likely number of successes (aka the mode) is given by:

Notice that they are the same!

Example 13.3 So let's look at when they aren't the same. let's flip the coin a fifth time. So $n = 5$ and $p = \frac{1}{2}$ again.

In this case, the expected number of successes is:

So we expect roughly two and a half heads to happen. That is the expected value.

Obviously this doesn't make sense since coin flips are static! We should have whole numbers in this case. That is where the mode kicks

in and gives us an actual number. The most likely number of successes is given by:

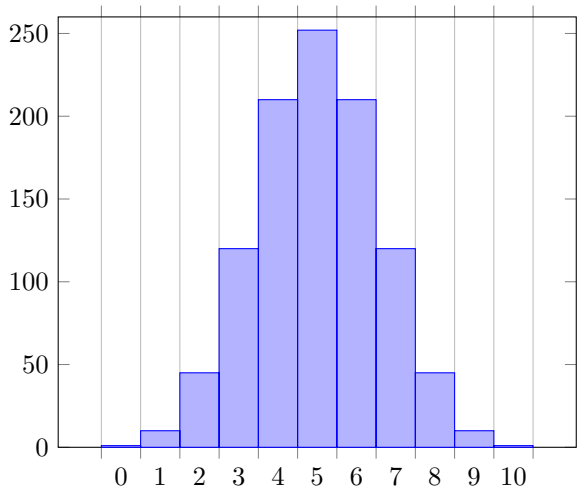
So we expect roughly 3 heads to occur.

This will be more discussed in more detail when we hit chapter three in the text.

14 Normal Distribution

Recall when we were looking at the binomial distribution we had the following example.

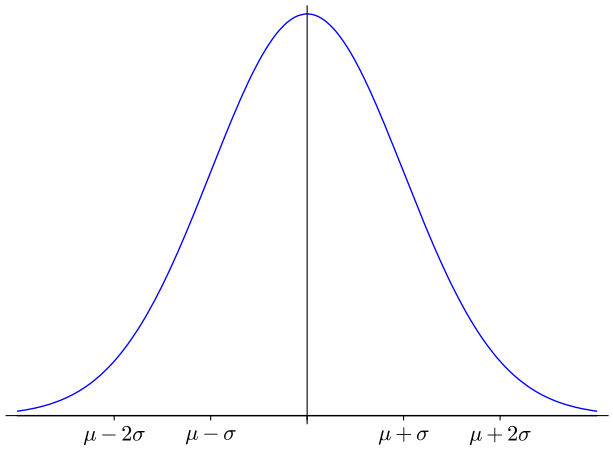
Example 14.1 The binomial distribution when $p = \frac{1}{2}$ and $n = 10$ is given by:



If you notice, we can kind of create a curve from these bars. This gives us what's called the normal curve.

The *normal curve* is the curve with equation:

The normal curve is often called the *bell curve* in the non-math world.



Five of these symbols you should know:

- x is a variable
- y is a variable (sometimes represented as $f(x)$).
- π is a constant roughly equal to $3.14159265358979\ldots$
- e is a constant roughly equal to $2.7182818285\ldots$
- μ is the *expected value*.

Wikipedia: Standard deviation

The only variable we haven't discussed so far is σ which is known as the *standard deviation*. Note that μ can be *any* real number while σ must be a strictly positive real number.

"Infinitesimally small width" should make you think of integration/calculus.

The best way to think about the normal curve is to think about it like a continuous histogram. In essence, it's when we start taking our boxes and making their with infinitesimally small. The μ tells us where the peak of the distribution is and the σ tells us how it's spread out.

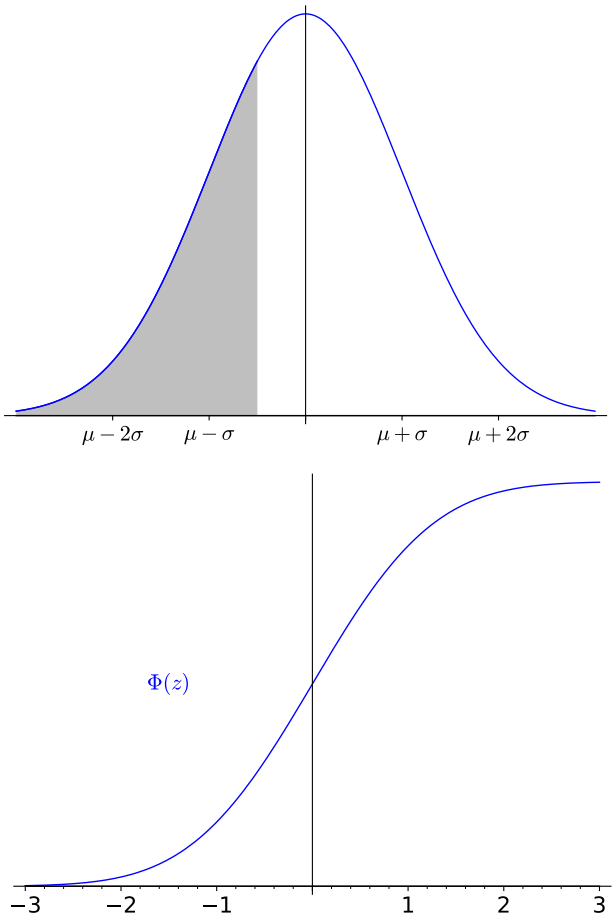
Note that the constant $\frac{1}{\sigma\sqrt{2\pi}}$ is only there to make sure that the area under the curve is equal to 1.

Wikipedia: Normal distribution
The normal distribution is sometimes called the "Gaussian distribution".

Although we have a curve, that doesn't actually define a distribution. Recall that for a distribution we need to define a probability map. The *normal distribution with expected value μ and standard deviation σ* is the distribution over the x-axis defined by the areas under the normal curve with μ and σ. If _____ then we call this distribution the *standard normal distribution*.

Wikipedia: Cumulative distribution function
We'll normally just right CDF to mean "cumulative distribution function".

Another way to look at the normal distribution is through a cumulative distribution function. For this we let $z = \frac{(x-\mu)}{\sigma}$ and we say that z is *x in standard units*. This terminology comes from the fact that in the standard normal distribution $z = x$. We let $\Phi(x)$ denote the probability of everything to the left of z. The function Φ is called the *standard normal cumulative distribution function*.



Since we're looking at area under the curve, we know that the standard normal CDF is given by the formula:

$$\Phi(z) =$$

The function _____ is known as the *standard normal distribution*. For a normal distribution the probability between a and b is given by:

If you recall from the end of your calculus class, $e^{-\frac{1}{2}x^2}$ doesn't have a simple indefinite integral, so $\Phi(z)$ has no simple exact formula unfortunately. But there are methods to calculate it.

Checkout Appendix 5 if you want some values just for fun.

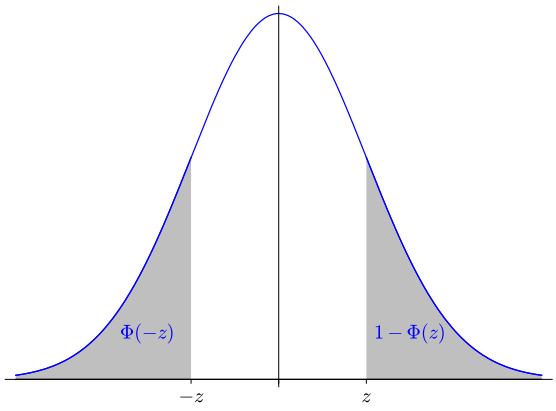
According to the book, there's a nice approximation function for $\Phi(z)$ given by:

$$\Phi(z) \approx 1-\frac{1}{2}\left(1+0.196854z+0.115194z^2+0.000344z^3+0.019527z^4\right)^{-4}$$

This approximation will give an approximation up to three significant figures for every $z \geq 0$.

Even without exact formulas there are some things we can deduce directly from the normal curve. For example, we know that

$$\Phi(-z) = 1 - \Phi(z) \text{ and } \Phi(0) = \frac{1}{2}$$



If we let $\Phi(a, b)$ denote the probability on the interval (a, b) then by the difference rule of probabilities we have:

We can also combine some formulas to get:

It's best not to memorize all of these formulas. It's better to try and know how we are deriving them. You can always look them up if you need to double check.

Plugging in some numbers we see:

$$\Phi(-1, 1) \approx \_\_\_\_\% \qquad \Phi(-2, 2) \approx \_\_\_\_\% \qquad \Phi(-3, 3) \approx \_\_\_\_\_\%$$

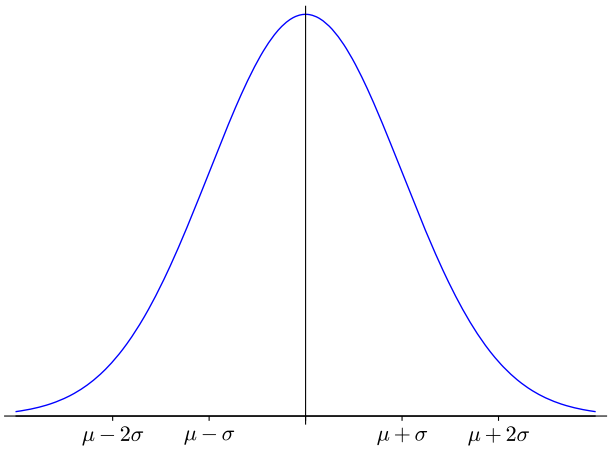
From these probabilities you can deduce almost any other probability.

Example 14.2 What is $\Phi(0, 1)$?

What is $\Phi(2, \infty)$?

15 The normal approximation to the binomial distribution

When we defined the normal curve, we used μ and σ, but we also mentioned that the idea came from looking at the curve of the binomial distribution.



Since the binomial distribution is dependent on n and p, we might ask how are the two pairs of numbers related to one another? One of them, μ, we already know. The expected value μ is just equal to ____. But what about σ? We won't go into details now, but it turns out that for sufficiently large n then $\sigma =$ _____. Remember that we called this σ the *standard deviation*.

Recall that in the normal curve, it doesn't make sense to ask about the probability at a certain point. So instead, let $P(a \text{ to } b)$ be the probability of getting i successes where $i \in [a, b]$ in n independent trials with probability p of success. Then we have that

$$P(a \text{ to } b) \text{ is}$$

$$\approx$$

$$=$$

In other words, the *normal approximation to the binomial distribution* for n independent trials with probability p of success is given by:

Let's see how this helps.

Example 15.1 Suppose that I randomly toss a fair coin $n = 100$ times. We want to know the probability of getting 50 heads.

By the binomial distribution we have:

By the numerical approximation we have that $a = b = 50$ and $\mu = np = 100 \cdot \frac{1}{2} = 50$.

Fluctuations If we make n grow more and more, we notice that our box sizes get smaller and smaller. This means that our approximations get better and better as n gets bigger. The size of the random fluctuations is of the order of $\sigma = \sqrt{np(1-p)}$ as it tells us how far from μ we are. So as $n \to \infty$ our σ gets larger and more of the distribution is covered.

Likewise, the proportion of successes fluctuates over time too. The typical size of random fluctuations in the relative frequency of successes is of order $\frac{\sqrt{p(1-p)n}}{n} = \sqrt{p(1-p)/n}$. Since $p(1-p)$ is maximized when $p = (1-p) = \frac{1}{2}$, we know $\sqrt{p(1-p)/n} \leq \frac{1}{\sqrt{4n}} = \frac{1}{2\sqrt{n}}$. And as $n \to \infty$ we get that $\frac{1}{\sqrt{n}} \to 0$ making it so that this order goes to 0 as n grows.

Theorem 15.2 (Square root law) *In n independent trials with n sufficiently large and probability p of success then*

- *The number of success will (with high probability) lie in a relatively small interval of numbers, centred on $\mu = np$, with width a moderate multiple of $\sqrt{n}$.*
- *The proportion of successes will (with high probability) lie in a small interval centred on p, with width a moderate multiple of $\frac{1}{\sqrt{n}}$.*

This implies one of my favourite theorems in probability:

Wikipedia: Law of large numbers

Theorem 15.3 (Law of large numbers) *Suppose that n is the number of independent trials and that n is very large. Let p be the probability of success on each trial. Then for every $\varepsilon > 0$*

$$P(\text{ proportion of successes in } n \text{ trials differs from } p \text{ by less than } \varepsilon) \to 1$$

as $n \to \infty$.

What this is saying is that the more trials we conduct, the closer to p our relative frequency will become.

Note that this is sometimes known as the "strong law of large numbers".

Week 4

3 June 2021

16 Confidence intervals

This idea of approximating the binomial distribution through the normal curve brings up an important topic that we'll discuss now. If we're doing a series of independent trials and we're recording how often we get successes, how do we know how close our rate is to the actual probability of success? For example, if I roll a die three times and I get 6 once, how do I know that 6 doesn't appear $\frac{1}{3}$ of the time? What if I told you it's not a fair die?

This is where we can take the ideas from the law of large numbers and use them to our advantage. Remember how the higher our n is, the more likely that our relative frequency was going towards our unknown probability. So if we had performed the die roll three trillion times and one trillion times we had a 6 show up, then we'd probably state that it was biased towards 6 at $\frac{1}{3}$ the time. Letting $\_\_$ denote our relative frequency, we know as n increases $\_\_\_\_\_$, but how can we measure it?

We can actually use the normal approximation. As we increased z we saw that $\Phi(-z, z)$ got closer and closer to 1. So for example, if $z = 4$ then $\_\_\_\_\_$ which means that if n is large enough, we can be 99.99% certain that the number of successes $n\hat{p}$ differs from np by less than $4\sqrt{np(1-p)}$. This means that the relative frequency $\hat{p}$ will differ from p by $\_\_\_\_\_$. Recall that $\sqrt{p(1-p)/n} \leq 1/2\sqrt{n}$ and so $4\sqrt{p(1-p)/n}$ is at most $2/\sqrt{n}$. In other words p is somewhere in the interval $\_\_\_\_\_$. This interval is called a 99.99% *confidence interval for p*.

Wikipedia: Confidence interval (binomial)

This confidence level is *specific* to the binomial distribution. You can define confidence intervals for other distributions, but we won't cover that now.

Wikipedia: Confidence interval (general)

Example 16.1 Let's look at a couple of examples. Say I redo my rolling of a die example and I roll the die roughly one million times (so

n is very large). We note that 6 appears roughly $180,000$ times. Find a 99.99% confidence interval for the probability that the die rolls a six.

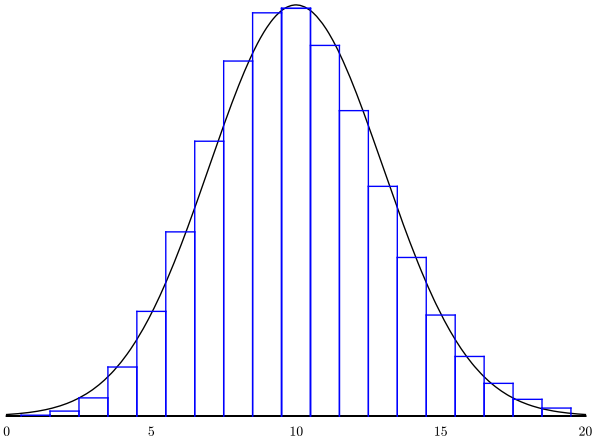
Notice that one of the key components to this approximation was the fact that $p(1-p) \leq \frac{1}{4}$. What this means is that the further from $\frac{1}{2}$ our probability p is, the less accurate this approximation becomes. Sure, we can just keep increasing n, but what happens if our probability is 1 in a bajillion? Then our approximation is not very good.

It turns out we can actually measure how good the normal approximation n is for a given n and p. Then let $N(a \text{ to } b)$ denote the normal approximation to a binomial probability $P(a \text{ to } b)$. Define $W(n,p)$ to be the *worst error* in the normal approximation to the binomial distribution, to be the biggest absolute difference between $P(a \text{ to } b)$ and $N(a \text{ to } b)$ over all integers a and b:

$$W(n,p) = \max_{0 \leq a \leq b \leq n} |P(a \text{ to } b) - N(a \text{ to } b)|$$

Running quick examples, you can check that for all $n \geq 10$ we know that $W\left(n, \frac{1}{2}\right) \leq 0.01$ and when $n \geq 20$ then $W\left(n, \frac{1}{2}\right) \leq 0.005$. This is why for $p = \frac{1}{2}$ this approximation is really good.

Now, suppose that $p \neq \frac{1}{2}$. Let's look at an example and see what happens and we might want to do to correct the problem. In the plot below, I set $n = 100$ and $p = \frac{1}{10}$.



What you'll notice is that the binomial distribution is slightly skewed to the left of the normal curve. This push to the left is known as the

skewness of the distribution. So what we will try and do is push the normal curve a little to the left in order to correct for this. We won't go to far into the details, but it turns out what we want to do is first take the third derivative of our function. Then

Wikipedia: Skewness

Wikipedia: Skew normal distribution

We then call the *skew-normal curve* the curve with the equation:

where $\text{skw}(n,p) = \frac{1-2p}{\sqrt{np(1-p)}} = \frac{1-2p}{\sigma}$.

Note that this skew function is *only* true for the binomial distribution. In general, it's a lot more complicated.

Notice that when $p = \frac{1}{2}$ then $\text{skw}(n,p) = 0$ giving us that $f(z) = \varphi(z)$. Depending on our choice of p then $\text{skw}(n,p)$ can either be positive or negative. It's positive if $p < \frac{1}{2}$ when the distribution is *skewed to the right*. It's negative if $p > \frac{1}{2}$ when the distribution is *skewed to the left*.

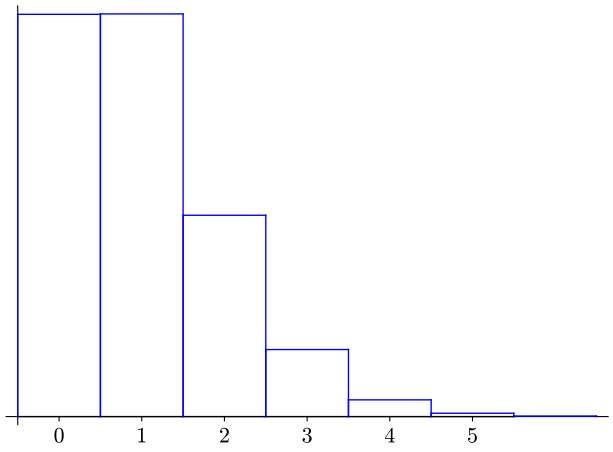
This gives us a second way to approximate the binomial distribution. For n independent trials where p is the probability of success, then

The term involving $\text{skw}(n,p)$ is known as the *skewness correction*.

Example 16.2 Let's use this approximation in an example. Suppose I want to calculate the distribution of 0s in 100 randomly selected digits. What is the probability of getting 4 or fewer 0s?

17 Poisson Approximation

Although normal approximation works nicely for probabilities near $p = \frac{1}{2}$, the further we get from 0.5, the worse the approximations will be. We tried to fix this by moving our approximation left and right, but even that has limits. The main issue is that the normal curve is symmetric (the left side and the right side look the same), whereas the binomial distribution is not a perfect curve. For example, we saw something like this earlier. If we let $n =$ _____ and $p =$ _____ we get the following distribution



It's so close to the 0 that there is no way we can make a nice bell curve that matches this. In this case we had let $p = \frac{1}{n}$ which means that the expected value is $\mu = np = \frac{n}{n} = 1$. If we follow this train of thought, as we do more and more trials, $n \to \infty$ and $p = \frac{1}{n} \to 0$, but μ stays at __. We can use this information to describe a curve which closely resembles the binomial distribution in this case. This "limit distribution" is called the Poisson distribution with parameter μ since we keep μ constant, but we take the limit as $n \to \infty$.

The *Poisson approximation* of the binomial distribution when n is large and p is small is given by

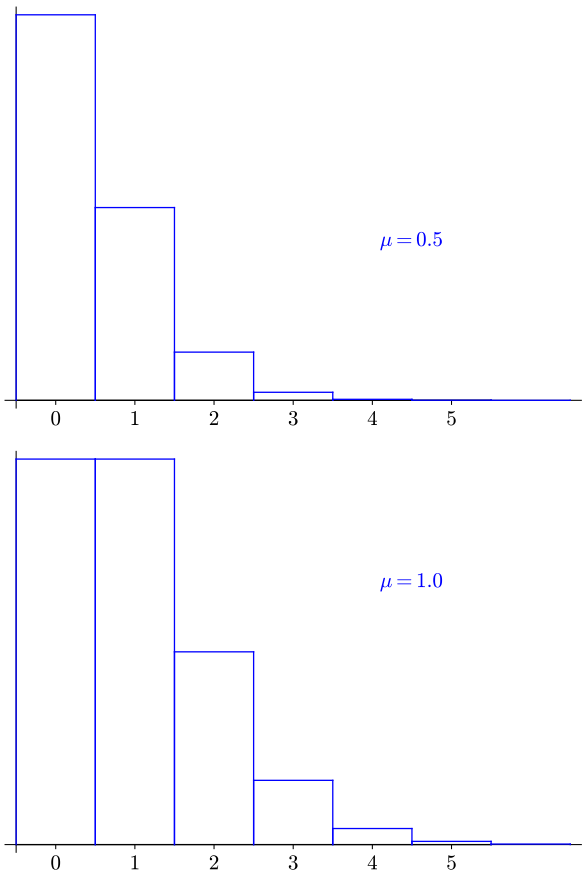
Just like with the normal approximation, the Poisson approximation is also itself a distribution. The *Poisson distribution* with parameter μ is the distribution given by:

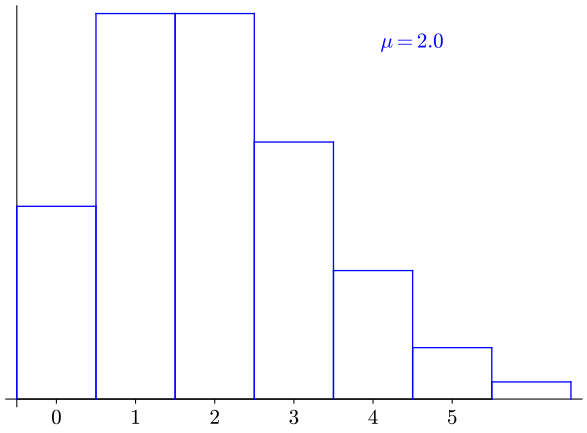
Wikipedia: Poisson distribution

Let's look at a couple examples.

Example 17.1 Every time you see a formula "$x - y$" there's a 1% chance you will copy it as "$x + y$". If you have to do 200 problems, what are the chances you will copy over 2 or more formulas incorrectly?

So what do these distributions look like as we change μ?





18 Random Sampling

Throughout the course so far we've seen random sampling various times without ever calling it random sampling. The idea of random sampling is to start off with a large population and to designate two categories of people. These categories might be something like "needs glasses" and "doesn't need glasses" or it might be "likes the colour red" and "dislikes the colour red". Basically the categories should be "opposite" of one another. We then take a small portion of the large population (a "random sample") and we look at the proportion of people that are in each category. This helps us figure out the proportion of people in each category for the population at large. Ideally, this would tell us exactly the proportion in the large population, but this is not always the case. Additionally, if we already know the proportion in the large population, we can ask what are the chances of a certain distribution happening in the smaller sample. We look at these questions through two main sampling methods

(1)

(2)

18.1 Sampling with replacement

Let N be the number of people in a large population and let n be the number of people in a sample which are drawn from the large population. These n individuals are drawn one at a time from the large population where each individual has the same chance of being chosen. After being drawn, they are put back into the large population (so a person *could* be chosen more than once). Therefore we would have a sequence of

length n of a set with N elements. This implies we have N^n different possible sequences which are all equally likely! Since we keep putting the individuals back into the large population, n *could* be larger than N.

An easy way to think about this is a bag full of N marbles. You take out a marble at a time and you record the colour and put the marble back into the bag. You do this n number of times. The question then becomes how well does the recording of the colours match with the actual distribution of colours in the bag of marbles?

Say R of the marbles are red and the rest B are not red (suppose they're blue just to make notation easier). Then _____. The distribution of red marbles is given by the probability $p = \frac{R}{N}$ which is the number we're trying to find. Notice here that we are working with a binomial distribution! Since p is static, the _____ approximation is a good approximation for the binomial distribution with parameters $\mu = np$ and $\sigma = \sqrt{npq}$. By the _____, if n is large enough then our sample of n marbles is highly likely to give the correct proportion of red to blue marbles. If you recall, from the section on confidence intervals, if we want a confidence interval of 99.99% we need to look at the interval: $\hat{p} \pm \frac{2}{\sqrt{n}}$ where $\hat{p}$ is the observed probability in our sample. If we only want a 95% accuracy, we can just look at the interval $\hat{p} \pm \frac{1}{\sqrt{n}}$.

Example 18.1 Say we have a bag with 20 marbles and we want to know how many are red. We pull out a marble, look at it's colour and put it back in. After doing this 35 times we notice that we have counted 25 red. What is the likely probability of a marble being red in the bag of 20 marbles? How about if we only pulled out 10 marbles and noticed 7 are red?

Recall that the other question we might want to ask is, if we already know the proportion in the large population, what is the probability of getting a certain distribution in the smaller sample? In this case, this should be fairly easy to calculate! If we already know that the probability of getting a red is $\frac{1}{p}$ then we have:

18.2 Sampling without replacement

Let's set things up the same as before, N individuals and we pull out n samples, but this time, we don't put the samples back in. Once we've taken someone out of the large population, we leave it out. This is just an ordering of n elements out of N individuals. So that means we'll have $(N)_n =$ _____ different options, which will be much less than our previous way of doing things.

In this case, we'll only look at the second question: if we already know the probability in the large population, what is the chances of getting the probability in n pulls?

So say that we have N total marbles and R of them are red and $B = N - R$ are blue. Then we know that the probability of red is given by $\frac{R}{N}$. Now we ask, say I pull out n marbles (without replacement), what is the probability that r of them are red?

First let's look at if the first r we pull are red and the rest $n - r$ are blue. Then we have:

for the chances of pulling a red out each time. Then, the rest of them are blue and so we need to multiply the above by:

Multiplying these two together we have:

But this is just for *one* option. We need to actually look at all combinations of r red appearing in n pulls. That means, we need to multiply the above by $\binom{n}{r}$ different combinations.

So in total we have:

19 Random Variables

We've been working on just one distribution so far: the binomial distribution. But if you were to look through Wikipedia, you'd notice that there are a ton of different distributions out there! The goal is to now generalize everything we have done into a more general setting. We start off with something that has been awkwardly missing in a math class: variables.

So far when we've talked about events, we've talked about them as subsets A of a sample space Ω. This works most of the time, but it can be hard to know what A is. For example, a lot of times we wrote $P(3)$ or something to mean "the probability of getting a 3". Instead of using the event "get a 3" which is represented by the subset $\{3\}$ we decided to just write the number 3. This is what we're going to try and generalize and we're going to create a whole new system for writing things out.

Variables have been used throughout mathematics as a placeholder for information, so it makes sense that we would use variables for our new system. We normally have these variables be large capital letters like X or Y, but they can be anything you prefer and they normally take the place of normal numbers. For example, if we are rolling a die, we might be interested in the event "the number 6 is rolled". To add a variable, we'd normally replace a number by a variable and go with that, *i.e.*, "the number X is rolled". Since we've been using capital letters for events, we'll let X be the variable we put into an event "the number X is rolled". In this case X is called a *random variable*.

Wikipedia: Random variable

As other examples, the random variable X might represent "the number X is rolled on a die" or "the side X of a coin is flipped" or "the

number X is drawn out of a hat", etc. Notice how in each case we don't actually say *which* number is rolled or *which* side is chosen, instead we let the random variable represent it.

How does this help? Well, let's look at the event from before represented by the random variable X: "the number X is rolled". We can let $X = 6$ mean _____. Notice how it gives us language to create new events. So if I were to say $X = 3$ then you'd know that we mean _____ in a much more compact way. We can then place these random variables into our probability function like _____ in order to mean _____.

At this point you might be asking, "why not just do something like _____" like before? Why are we complicating things? Because complicating things sometimes makes things less complicated! For example, what about if I want to ask for the event "the number rolled is less than or equal to 3"? Now, we can just write _____ and we know exactly what we're talking about! Even with more complicated events like "Let A be the subset $\{2, 4, 6\}$ of all even numbers and find $P(A)$", we can do this easier with our new random variables. We can represent the above as _____.

A random variable doesn't always have to represent a number! If we look at flipping a coin, I can ask for $P(X = H)$ where H represents getting a heads. These random variables can mean any particular outcome in our sample space!

Here's a quick table of everything from above to maybe help. (Suppose we are rolling a fair six-sided die)

English language	Random Var.	Subset	Probability
Number on the die is 6	$X = 6$	$\{6\}$	$\frac{1}{6}$
Number on the die is less than 3	$X < 3$	$\{1, 2\}$	$\frac{1}{3}$
Number on the die is x	$X = x$	$\{x\}$	$\frac{1}{6}$
Number on the die is less than x	$X < x$	$\{1, 2, \ldots, x-1\}$	$\frac{x-1}{6}$
Number on the die is in the subset B	$X \in B$	B	$\frac{\lvert B\rvert}{6}$

If we use X to help define an event A, then we say that the *event A is determined by X*. Although we *can* write $P(A)$ to let us know the probability, we will henceforth start writing $P(X \in A)$ to show that A has some variable in it. If we go over all possible subsets A, we (must) get a distribution which we call the *distribution of X*. In essence, the outcome of any particular outcome x is given by _____ and of any subset, via the addition rule, by _____

At this point, it might confusing and you might ask what the difference between x and X are. This is super confusing, so don't worry! A random variable is just a normal variable that represents certain things

inside of events. When we talk about the probability of an event then we must state what the random variable is equal to for the probability to make sense. We can have the random variable be equal to a standard variable (_____), but we can also have it mean anything else! One thing that never makes sense is _____ as this doesn't tell us anything.

The book uses the term "dummy variable" to mean a standard-/normal variable.

So say I have an event"I roll a five sided die and a I get an X" and an event "I pull the number Y out of a hat". Then if were to say _____ then what we mean is that the probability that we get u when rolling a five sided die is the same as the probability of pulling out v from out of a hat.

Aside: As an aside, if you look at the definition of a random variable in Wikipedia, it might be a little confusing since we're not defining it in exactly the same way. Since in mathematics we like to be precise, the exact definition of a random variable is given through a certain type of function. So X would be $X : \Omega \to E$ where E is some space (normally $\mathbb{R}$ for us). Then, more precisely we have $P(X \in A) = P(\{x \in \Omega \mid X(x) \in A\})$. Don't worry about this to much, but it's a good thing to know.

Example 19.1 Let's look at a quick example of what this looks like. Say I roll a fair six-sided die and I want to calculate some probability functions.

$$P(X = 2) =$$

$$P(X \leq 4) =$$

$$P\left(\frac{X}{2} \in \mathbb{Z}\right) =$$

20 Functions

Sometimes we want to look at random variables as a function of another random variable. Normally we see this in standard variables by: $y = f(x)$. Doing this for random variables we get: $Y = f(X)$. What does this mean though?

It means that if X has some value then Y has the value $f(X)$. The two are related to one another through the function. This implies that the distribution of Y can be derived from the distribution of X.

Example 20.1 Suppose we're rolling two dice and we want to calculate the sum of the two values. We've done this many times, but we're going to be using functions this time.

As before, we let (i, j) represent a roll of the two dice. If we let X be the sum of our dice, then we know that we get a distribution like the

following:

x	2	3	4	5	6	7	8	9	10	11	12
$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

These were calculated from the (i,j). Let Y represent the outcomes (i,j) and let f be the function $i+j$. In other words _____.

This is obviously a super easy pointless example. We're not really gaining any meaningful insights. What we *do* learn though is that we can use functions in conjunction with our variables. Addition doesn't do much, but there are limitless possibilities!

21 Joint Distributions

Let's take this one step further. What happens if an event has potentially more than one variable? For example, "What is the probability that I get the 5 of spades in a normal deck of cards?" I could technically represent this as one variable: "What is the probability that I get a X in a normal deck of cards?". But what if I want to ask for the probability of getting a 5, regardless of suit? I can no longer ask that! In this case it makes more sense to put in two variables: "What is the probability that I get a X of Y in a normal deck of cards?". Now we can ask actually ask for the probability! We would get _____ where S is the set of suits. How about our original example? This we can write as _____.

Wikipedia: Joint distribution

Whenever we have two or more random variables in a distribution we call it a *joint distribution*. Joint distributions can be a little weird to work with, but once you've done a few examples, they make more sense.

We'll look at two examples to try and make things more understandable.

Example 21.1 First we'll look at an example of pulling marbles out of two bags. Say that each bag has 3 red marbles and 1 blue marble. Our event is represented by "What is the probability of pulling out a X

out of bag 1 and a Y out of bag 2". Let's see what the chances are using a table.

	X =Red	X =Blue	Total
Y =Red	$\frac{3}{4}\cdot\frac{3}{4}=\frac{9}{16}$	$\frac{3}{4}\cdot\frac{1}{4}=\frac{3}{16}$	$\frac{12}{16}=\frac{3}{4}$
Y =Blue	$\frac{1}{4}\cdot\frac{3}{4}=\frac{3}{16}$	$\frac{1}{4}\cdot\frac{1}{4}=\frac{1}{16}$	$\frac{4}{16}=\frac{1}{4}$
Total	$\frac{3}{4}$	$\frac{1}{4}$	1

So we can read this table by looking at each column and row for each random variable. For example, $P(X=\text{Red}, Y=\text{Blue})$ gives us ____.

We can also technically look at any particular row and column to get the added probability. For example, if we want to know the probability that the first marble is red, then we're asking for X to be "Red". We can represent this as:

In particular, we can always use that equality whenever we have multiple random variables:

Example 21.2 Let's look at the previous example, but just look at one bag. So say we have one bag with 3 red marbles and 1 blue marbles and I pull out two marbles (one after another without replacement). Here I can represent this as the event "I pull out marble X then I pull out marble Y". So looking at the probability table we have:

	X =Red	X =Blue	Total
Y =Red			
Y =Blue			
Total			

We say that two random variables are *equal in distribution* if

The book calls this *same distribution*

If two random variables are equal in distribution, then we can change their variables whenever we want. For example _____ and we'll get the same thing!

We say that two random variables are *equal* if

Another way of saying this is:

Equal in distribution basically says the right column and the bottom row must be the same. Equal says the sum of the (main) diagonal must be 1

Notice how we can take this idea further:

$P(X < Y) =$

or even further:

$P(X+Y = z) =$

Example 21.3 Calculate the distribution of $X + Y$ if we roll two fair six-sided dice.

22 Conditional Distributions

Joint distributions might be confusing, but if you think about it, we had already been working with joint distributions for a while! Whenever we had a conditional probability statement $P(A \mid B)$, we were working with two different events/variables at the same time. The only issue was that it wasn't done in the framework of random variables. We now think about everything through the eyes of a random variable.

So let's now say we want to replace the event A by a random variable. We saw that $P(A)$ is written as _____ using random variables, so what would our conditional probability look like?

What if we now replace B with a random variable? We get:

Two random variables, easily put in! And just like before, we can switch these variables out to whatever we like. In particular, we'll define the *conditional distribution of X given $Y = y$* as

What happens when the events coming from X and Y are independent? What does independence mean in this case? We had this definition for independence $P(A \cap B) = P(A) \cdot P(B)$ but how do we move that over to joint distributions? We say that two random variables are *independent* if

Week 5

10 June 2021

23 Multinomial Distribution

What if we have many different events happening all at once? We can easily take our idea of joint distributions and expand it to multiple random variables.

Through this we can talk about a generalization of the binomial distribution. Remember that the binomial distribution had two potential outcomes "yes" and "no" with probability p and $(1-p)$. What happens if we allow multiple possible outcomes with their own probabilities?

Recall that for the binomial distribution what we did was we made a tree where at every node, the tree split into two parts with probability p and $(1-p)$ at each part. We knew the coefficient at any given node because it was given by the binomial coefficient $\binom{n}{k}$. So if we did n trials and wanted to see if k came back, then we looked at

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

To construct the multinomial distribution, we construct the same tree as before, but at every node we split the tree into ℓ different parts. If $\ell = 2$ this is called the binomial distribution and we saw that this tree is just Pascal's triangle. If $\ell = 3$ this is called the *trinomial distribution* and this tree is given by Pascal's pyramid. For an arbitrary tree, we would have ℓ edges coming out of each node and the probability on each edge would be labelled $p_1, p_2, \ldots, p_\ell$ where _____. If we then perform n total trials and wanted to see if k_1 of them were for the first option, k_2 for the second, etc. then we would get:

Wikipedia: Pascal's pyramid

Binomial distribution is the multinomial distribution with $\ell = 2$.

but what is our constant here? It turns out, it's something called the "multinomial coefficient" which come from the multinomial theorem:

Wikipedia: Multinomial theorem

which is basically what we have above! So if we plug in p_i for each x_i we get

Example 23.1 Say that we're rolling a dice and we keep track of three different variables:

(1) Even numbers

(2) The number one

(3) Everything else (aka the numbers three and five)

If we roll the dice 10 times what's the probability that we get 6 even numbers and the number one 3 times?

24 Symmetry

We now look at the histograms of a given distribution and we ask when is our probability distribution symmetric. You might remember the word "symmetric" through calculus where you discussed that an even function is symmetric about the y-axis (aka you can reflect it over the y-axis and get the same function) and that an odd function is symmetric about the origin (aka you can rotate it by half a circle around the origin and get the same function). For us, we'll say that a distribution of X is *symmetric about v* if

Wikipedia: Symmetric distribution

When $v = 0$ we get symmetry about the y-axis. Equivalently, we can look at things from an inequality perspective:

Now, say that we take the sum of n random variables:

$$S_n = X_1 + X_2 + \ldots + X_n$$

where each X_i is symmetric about 0, *i.e.*, _____
Then we know that X and $-X$ have the same distribution since:

If we're doing independent trials, we know that this implies that S_n and $-S_n = \sum -X_i$ have the same distribution. The way I think about this is: each X_i has the same distribution as $-X_i$; and since each random variable is not dependent on the others, the distribution of S_n (aka looking down the diagonal in our tables) is the same as $-S_n$ (looking down the diagonal, but with -1). In other words _____
Therefore S_n is symmetric about 0.

We can do a similar thing above if the (independent) X_i are symmetric about v_i. In this case, we must do a change of variable. We let $Y_i = X_i - v_i$ and in this case each Y_i is symmetric about 0. Letting $S'_n = Y_1 + Y_2 + \cdots + Y_n$ then from above we know $-S'_n$ and S'_n have the same distribution. Replacing the variables we have:

And so

Then

In other words S_n is symmetric about $\sum v_i$.

Example 24.1 Suppose that we take 101 (independent) random numbers from the set $\{0, 1, 2, \ldots, 9\}$. Find the probability that the sum of the numbers is less than 455.

25 Expected Value

We saw in the sections on the binomial distribution something called the "expected value" which most of you have probably seen as the mean since it's calculating the average probability. This was defined as $\mu = np$ where n was the number of trials which had a probability of p of success. This makes sense since normally when we think of the mean, we think

about the average value of something:

$$\frac{x_1 + x_2 + \ldots + x_n}{n}$$

We can actually think of this in a probabilistic sense and look at how many times each x_i occurs and group them together. For example:

We can rewrite this as

Wikipedia: Expected value

We can use this formula to calculate the "mean" for arbitrary distributions The *expected value* of a random variable X is

Example 25.1 Suppose that we roll a fair six-sided die randomly. What is the expected value?

25.1 Indicator function

If you think about it, the binomial distribution was kind of weird in that we weren't talking about the expected value of just *one* trial, we were asking the expected value over *multiple* trials. So let's see how this kind of works.

Let's look at a binomial distribution where we run n trials each with a probability of success p. If we focus on just one trial, what do we get?

First we need to focus on how to represent this. Since our trial either succeeds or fails we can create a random variable where $X = 1$ if we're successful and $X = 0$ if we're not. This is called an *indicator function* of an event since it *indicates* whether or not an event succeeded. Then

Wikipedia: Indicator function

the expected value of the indicator function is given by:

What happens if we run this trial n times now? What is the expected number of successes? So if we let X_i be the indicator function that the ith trial was successful, then what we're asking is after n trials, for how many X_i do we have $X_i = 1$. We can calculate this using the following expression:

This is known as *linearity for expectation*. Note that it *does not* depend on whether the random variables are independent or not! So what this means is that, in the binomial distribution case we have:

The book calls this linearity the "addition rule for expected value"

since each $E(X_i) = p$ as they are all independent and each X_i is successful with probability p.

The cool thing about the addition rule is that we can alter our probabilities and look at the success of multiple events! So for example, if I have n different events and I let X_i be the indicator function for each event, *i.e.*, $X_i = 1$ if the ith event is successful. Then let p_i be the probability that the ith event is successful. By everything above, we have:

Example 25.2 Suppose that we do a binomial distribution where we draw a card 20 times with replacement, always looking for the ace of hearts. What is the expected number of successes?

Example 25.3 Suppose that we draw 5 cards from a deck of 52 cards and we end up with X number of twos. What is the expected number of twos among the five cards?

Week 6

17 June 2021

26 Tail formula

Last time we talked about the expected value of a distribution:

We saw that, if we're just trying to figure out whether something succeeded or not, we can look at indicator functions instead of the probability itself. In this case, we let X_i denote the indicator function and we end up with:

Since this equation looks so nice, we're going to see whether we can turn our original definition of $E(X)$ into something nicer. It turns out, we can in the case where our x are coming from a set $\{0, 1, \ldots, n\}$. In this case, if we calculate $E(X)$ we get:

$$\begin{aligned} E(X) =& P(X = 1) \\ &+ P(X = 2) + P(X = 2) \\ &+ P(X = 3) + P(X = 3) + P(X = 3) \\ &+ P(X = 4) + P(X = 4) + P(X = 4) + P(X = 4) \\ &+ \cdots \end{aligned}$$

We can rewrite this into what's known as the *tail sum formula for expected value*.

Example 26.1 Suppose that we roll four fair six-sided dice and we let M be the minimum value of the four numbers rolled. What is the expected value of M?

27 Markov's Inequality

Let's go back to indicators again. For indicators, we took the sum of the indicators to get the expected value of success of our event X. How do we know if our sum of indicators is itself an indicator?

So let $X_1, X_2, \ldots, X_n$ be a bunch of events and let $I_1, I_2, \ldots, I_n$ be

their indicator functions. In other words, if X_i occurs then $I_i = 1$, if X_i does not occur then $I_i = 0$. If we think about it, an indicator only has values 0 and 1; that is the definition of an indicator function. So what about if we're asking for the sum of indicators?

$$I = I_1 + I_2 + \cdots + I_n$$

When is I an indicator? Well, I is an indicator if and only if I is equal to 0 or 1 (by definition). Since each of the I_i can only have values 0 or 1, that means *at most* one of the indicator functions can be 1 at any given calculation. But this is true if and only if the events X_i are mutually exclusive, *i.e.*, no two events can happen at the same time. In this case, we get the following equality:

If the events X_i are *not mutually exclusive*, then we get what's known as *Boole's inequality*.

Wikipedia: Boole's inequality

We can notice a couple things from this inequality by looking at the indicator functions. Let $I = I_1 + \cdots + I_n$ as before, then the right hand side is equal to $E(I)$. The left hand side, we can view it as the probability that at least one thing is true. In other words $P(I \geq 1)$. So we can rewrite the above as:

We can generalize this for arbitrary expected value to get what's called *Markov's inequality*.

Wikipedia: Markov's inequality

Example 27.1 Suppose we have a (non-negative) random variable X and we know its expected value is equal to 5. What is the largest probability of success that $P(X \geq 100)$ could be?

28 Multiplication Moments

We've talked about how we can make random variables functions of one another. The example we gave wasn't that great, *but* we'll be using functions to see how we can easily calculate the expected value from one random variable into another.

Let $Y = f(X)$. Then

$$E(Y) =$$

There are particular functions which mathematicians like to look at in particular. Let $f(x) = x^k$ for some strictly positive integer k. Then

$$E(X^k) =$$

Wikipedia: Moment

The distribution X^k is known as the *kth moment of X*.

If $k = 1$ then we have the *first moment* which is usually called the expected value (or mean). If $k = 2$ then it's called the *second moment* or, alternatively, the *mean square*.

Example 28.1 Let's calculate the mean square of the discrete uniform distribution of the set $\{0, 1, 2\}$. First, notice that

$$E(X) =$$

The mean square is then given by:

$$E(X^2) =$$

Notice how in general $E(f(X)) \neq f(E(X))$. In other words

Let's use functions in order to see why $E(X + Y) = E(X) + E(Y)$. Suppose we have two random variables X and Y and we want to calculate $E(X + Y)$. To do this, let's make a function f which takes in two random variables and outputs a third: $f(X, Y) = X + Y$. By the above we have

Another thing to note is that if we multiply a random variable by a constant, then we can just pull the constant out:

$E(cX) =$

What happens when we try and multiply two random variables?

But we can't actually separate here! The only time we can actually separate is if X and Y are independent. In that case:

In other words, if X and Y are independent, then

29 Variance

When we were working on the binomial distribution, we had defined two new concepts: the expected value and standard deviation. We already handled the expected value, so let's move onto the standard deviation. We had that the standard deviation was defined as $\sqrt{np(1-p)}$ where np was the expected value.

What was the standard deviation keeping track of? It was keeping track of how far away from the expected value we were. If we look at what's inside the square root and we expand it, we get $np - npp = E(X) - E(X)p$. In other words, we get the expected value and then a little correction factor.

If we let $\mu = E(X)$ be our correction factor, then

Since we'll eventually want to take the square root, we avoid having a negative number by squaring our equation. This gives us what is called the *variance* of a distribution

Wikipedia: Variance

Wikipedia: Standard Deviation

The *standard deviation* is then the square root of the variance.

$$\mathrm{SD}(X) =$$

Example 29.1 Suppose we have a fair eight sided die. Find $\mathrm{SD}(X)$ where X is the number on the die after one roll.

What if we only cared about success/failure of an event? In this case, we look at indicator functions and the formulas above become much easier. Remember that whenever we have an event A, we can create a random variable X_A that indicates whether or not A succeeded. In other words: X_A is either equal to 1 or 0 and $E(X_A) = P(A)$. Then

since X_A is equal to 1 or 0 we have:

$$X_A^2 : \ 1^2 = 1 \quad 0^2 = 1 \ \Rightarrow \ X_A^2 = X_A$$

So we have that

$$\text{Var}(X_A) = E(X_A^2) - E(X_A)^2 = E(X_A) - E(X_A)^2 = P(A) - P(A)^2 = P(A)\left(1 - P(A)\right)$$

30 Standardization

If we have constants a and b then $E(aX + b) =$ _____.
Plugging this into the standard deviation we get:

We can use these formulas to create a standardized version of the expected value and the standard deviation. Let $X^\star = \frac{X - E(X)}{\text{SD}(X)}$ Then:

$E(X^\star) =$

and

$\text{SD}(X^\star) =$

These are called the *standard units* for the distribution X.

This should look super familiar! Let $\mu = E(X)$ and $\sigma = \text{SD}(X)$, then the fractions we're looking at are _____.

Example 30.1 The average height for men in Canada is 178 cm and the standard deviation is roughly 7.5 cm. Approximately what percentage of Canadian men are taller than 166 cm?

For women, the average Canadian is 164 cm with a standard deviation of 7 cm. Unfortunately there is no data on non-binary or intersex individuals.

31 Chebychev's Inequality

Since the standard deviation tells us how far away things will stretch from the expected value, there must be some relation between the two. It turns out that given a random variable, the probability that it differs from its expected value by more than k standard deviations is at most $\frac{1}{k^2}$. This is known as *Chebychev's inequality*:

Wikipedia: Chebyshev's inequality

Example 31.1 Say we're working for a financial company who are looking at all the transactions of their customers. They notice that given one billion transactions, the average transaction is roughly \$20 and if we square the values of the transactions, the average becomes \$404. Find an upper bound on how many transactions are over \$50.

32 Central Limit Theorem

Before we get into some laws and theorems, we're quickly going to state what variance looks like under addition. If X_1, X_2, ..., X_n are all mutually independent, then

This doesn't work for dependent variables unfortunately! For example, if $X = Y$ (so they are dependent on one another) then we have

$$\operatorname{Var}(X+Y) = \operatorname{Var}(2X) = \operatorname{SD}(2X)^2 = (2\operatorname{SD}(X))^2 = 4\operatorname{Var}(X)$$

and

$$\operatorname{Var}(X) + \operatorname{Var}(Y) = \operatorname{Var}(X) + \operatorname{Var}(X) = 2\operatorname{Var}(X)$$

Remember how for the binomial distribution we had something called the square root law where we basically stated that as we increase the number of trials then most trials would be close to the expected value. This works with any distribution. We'll go through this slowly.

First let's suppose we have n independent random variables X_i each with the same distribution X. In other words _____ for all i. Since expectation and variance are determined by distributions, we also know $E(X_i) =$ _____ and $\operatorname{Var}(X_i) =$ _____. We

let S_n be the sum of all the random variables:

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

Then:

$E(S_n) =$ $\quad$ $\mathrm{Var}(S_n) =$ $\quad$ $\mathrm{SD}(S_n) =$

This gives us the *square root law*:

Theorem 32.1 (Square root law) *Let S_n be the sum of n independent random variables $X_1, X_2, \ldots, X_n$ each with the same distribution X. Let $\bar{X}_n = \frac{S_n}{n}$ be the average value. Then*

We continue our trek by looking at the law of large numbers in this context. As n increases, we see that $\mathrm{SD}(S_n)$ will grow while $\mathrm{SD}(\bar{X}_n)$ decreases. This simple idea gives us the law of averages.

Wikipedia: Law of averages

Theorem 32.2 (Law of averages) *Let S_n be the sum of n independent random variables $X_1, X_2, \ldots, X_n$ each with the same distribution X. Let $\bar{X}_n = \frac{S_n}{n}$ be the average value. Then for every $\varepsilon > 0$*

Notice that we don't have an approximation for S_n in the theorem above. That's because there is no simple formula for the distribution of S_n. Instead, we can use normal approximation to find a simple approximation for S_n.

Wikipedia: Central limit theorem

Theorem 32.3 (Central limit theorem) *Let S_n be the sum of n independent random variables $X_1, X_2, \ldots, X_n$ each with the same distribution X. For large n, the distribution of S_n is approximately normal, i.e., $E(S_n) = nE(X)$ and $\mathrm{SD}(S_n) = \sqrt{n}\,\mathrm{SD}(X)$. In other words:*

Where Φ is the standard normal CDF.

Wikipedia: Random walk

Example 32.4 In this example, we'll consider what's known as a random walk. The problem is normally told from a physics perspective as

this is where the idea came from.

Suppose you have an infinite line of slots and you put a particle in the middle. Each second, the particle moves left with probability p_ℓ, right with probability p_r or stays where it is with probability p_s, *i.e.*, $P_\ell + p_r + p_s = 1$. Let's suppose that the particle is having a lazy day. It stays where it is half the time $p_s = \frac{1}{2}$ and it moves with equal probability, *i.e.*, $p_\ell = p_r = \frac{1}{4}$.

There are roughly $86,400$ seconds in a day, so let's say we run this experiment for a little longer than a day. After $88,200$ seconds, what are the chances that the particle is 300 slots away to the right from where it began.

33 Skewness

Just like with the binomial distribution, sometimes our normal approximation isn't very good. We need to add some sort of correcting factor. Since in the central limit theorem we look at $\frac{S_n - E(S_n)}{\text{SD}(S_n)}$ we let $X_\star = \frac{X - E(X)}{\text{SD}(X)}$ and look at our approximations from this perspective. In this case, the first moment is given by:

$E(X_\star) =$

and the second moment is given by:

$E(X_\star^2) =$

In order to find how much we're off by, we look at the third moment, and define skewness in that way.

$$\text{skw}(X) = E(X_\star^3) =$$

If $S_n = X_1 + X_2 + \ldots + X_n$ where the X_i are independent each with the same distribution X, then

$$\text{skw}(S_n) =$$

We won't try and show where these formulas come because they are difficult to show. If you do want to try and prove it you can do it by first showing:

$$E\left((S_n - E(S_n))^3\right) = nE\left((X - E(X))^3\right)$$

and

$$\mathrm{SD}(S_n) = \sqrt{n}\,\mathrm{SD}(X)$$

Putting these together gives you the above formula.

As a quick example when $n = 2$ we have the following:

$$\begin{aligned} E\left((X_1 + X_2 - E(X_1 + X_2))^3\right) &= E((X_1 + X_2)^3) + 3E((X_1 + X_2)^2)E(X_1 + X_2) \\ &\quad + 3E(X_1 + X_2)E(X_1 + X_2)^2 + E(X_1 + X_2)^3 \\ &= E(X_1^3) + 3E(X_1^2)E(X_1) + 3E(X_1)E(X_1)^2 + E(X_1)^3 \\ &\quad + E(X_2^3) + 3E(X_2^2)E(X_2) + 3E(X_2)E(X_2)^2 + E(X_2)^3 \\ &= 2E((X - E(X))^3) \end{aligned}$$

and

$$\begin{aligned} \mathrm{SD}(X_1 + X_2) &= \sqrt{\mathrm{Var}(X_1 + X_2)} \\ &= \sqrt{\mathrm{Var}(X_1) + \mathrm{Var}(X_2)} \\ &= \sqrt{2}\sqrt{\mathrm{Var}(X)} \\ &= \sqrt{2}\,\mathrm{SD}\,X \end{aligned}$$

Week 7

1 July 2021

34 Geometric Distribution

We start off this week with a question which was posed in chapter 1 of the book, but I waited until now to talk about. Say you're rolling a die multiple times and you want to see how many times you have to roll it until you get a 6. The first roll, you have a $p = \frac{1}{6}$ chance of getting a six. That means you have a _____ chance to roll a six in 1 roll and a _____ chance to roll a six in 2 or more rolls.

The chances of a six showing up on the second roll is: $\frac{5}{6} \cdot \frac{1}{6}$ where the $\frac{5}{6}$ is coming from you *not* rolling a six in the first roll. So to roll a six in the first two rolls gives us $\frac{1}{6} + \frac{5}{6}\frac{1}{6} = \frac{11}{36} \approx 0.3056$. Therefore, there's a _____ chance of rolling a six in your first two rolls.

What about three? Well, we do the same. We want to add a third roll to our list. In this case, we need to fail twice and succeed the third time so we get: _____. This gives us the probability of rolling a six only on the third roll. Adding this to our second sum gives:

$$\frac{1}{6} + \frac{5}{6} \cdot \frac{1}{6} + \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \frac{91}{216} \approx 0.4213$$

which is the probability we roll a six in our first three rolls.

Notice how we're getting a little pattern here. For k rolls, we have

If we replace $\frac{1}{6}$ with p and $\frac{5}{6}$ with $(1-p)$, we end up with _____ But, with a little trick, we can actually simplify this. Let $q = 1 - p$ and

then we get that $p = 1 - q$. Plugging this in we have:

$$(1-q)\sum_{i=1}^{k} q^{i-1}$$

In other words, the chance of rolling a six in the first 3 rolls is:

With our notation we also have that the chance of rolling a six *on* the fifth roll is _____. Generalizing this gives us a probability distribution which is known as the *geometric distribution*. Let X be the number of times that something happens with probability p and let $q = 1 - p$.

Wikipedia: Geometric distribution

35 Discrete Distribution

Notice how for the geometric distribution, we can let k be any number! In other words, our sample space is no longer finite. This is the next step we take in our study of distributions from a discrete angle. A *discrete distribution* is a probability distribution where the sample space is the set of non-negative integers $\{0, 1, 2, 3, \ldots\}$ with a sequence of probabilities $p_0, p_1, p_2, p_3, \ldots$ such that each $p_i \geq 0$ and $\sum_i p_i = 1$.

Wikipedia: Discrete distribution

Examples of discrete distributions are:

- Discrete uniform distribution
- Bernoulli distribution
- Binomial distribution
- Geometric distribution
- Poisson distribution

Note that we do need to introduce one rule in order to make discrete distributions a distribution.

Theorem 35.1 (Infinite sum rule) *If an event A is partitioned into $A_1, A_2, A_3, \ldots$:*

$$A_1 \sqcup A_2 \sqcup A_3 \sqcup \cdots$$

then

$$P(A) = P(A_1) + P(A_2) + P(A_3) + \ldots$$

With that, all of our rules are satisfied for discrete distributions to be distributions.

Note that finite distributions are discrete distributions! For example, if we want to just roll a dice the distribution involves only six numbers: 1, 2, 3, 4, 5, 6. Where

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

In order to view this as a discrete distribution, you just need to let $P(X=i)=0$ whenever $i \notin \{1, 2, 3, 4, 5, 6\}$.

Example 35.2 Let's say that your friend decides to tell you that they're luckier than you. They pick up a dice and say, "I bet you I can get a six before you can". Before you agree, you decide to calculate the chances that they're right. What's the probability that your friend will win?

36 Discrete moments

What can we say about the expected value and, more specifically, the moments of a discrete distribution. The expected value comes naturally

$$E(X) = \sum_x xP(X = x)$$

but only if this sum is well-defined! So, we say that the expected value is $E(X)$ (as defined above) *if* the series is absolutely convergent:

Why absolute convergence and not normal convergence? This comes from the Riemann series theorem which states that we can only rearrange the terms in our series if the series is absolutely convergent. Note that all of our discrete distributions *are* convergent if and only if they are absolutely convergent. Why? Because we stated that $P(X = x) = p_x \geq 0$ and we said that our sample space is all non-negative integers! So a non-negative integer times a non-negative integer is always non-negative! We only need to care about absolute convergence if we allow negative terms in our sample space.

What if we look at an arbitrary function instead of just X for the expected value? Things like the moments $E(X^2)$, $E(X^3)$ or just any other function. Supposing that our function is $f(X)$ then

$$E\left(f(X)\right) = \sum_x f(x)P(X = x)$$

is defined if the right-hand side is absolutely convergent.

Example 36.1 Let's go back to the geometric distribution we've been working with most of this week. What is the expected value for when the first six will show up when rolling a fair six-sided die? In other words, how many times do we expect to roll the die before the first six shows up?

Calculate the standard deviation.

Example 36.2 Let's look at a second example since these things can get a little confusing. Say that McDonald's has a new collection of toys for their happy meals and you've decided you want every single toy! Suppose that the chance of getting each toy is independent and there are n toys in total. If each purchase gives you a toy at random, how many happy meals are you expected to buy before you get every toy?

37 Poisson Distribution

We're now gonna take a step back and look at a distribution from way back in the day: the Poisson distribution. In Week 3 we had talked about the binomial distribution where we have some trial which is successful with probability p. We then repeat this trial n times and we ask what is the probability that exactly X were successful. There were two ways we could approximate this: the normal distribution for when p is close to $\frac{1}{2}$ and the Poisson distribution for when p is sufficiently small. Recall that the Poisson approximation was given by:

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

where $\mu = E(X) = np$.

The *Poisson distribution with parameter μ* is the distribution: Wikipedia: Poisson distribution

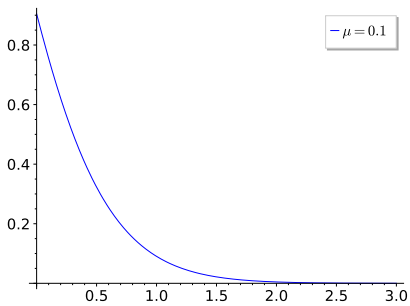
$$P(X=k) = e^{-\mu}\frac{\mu^k}{k!}$$

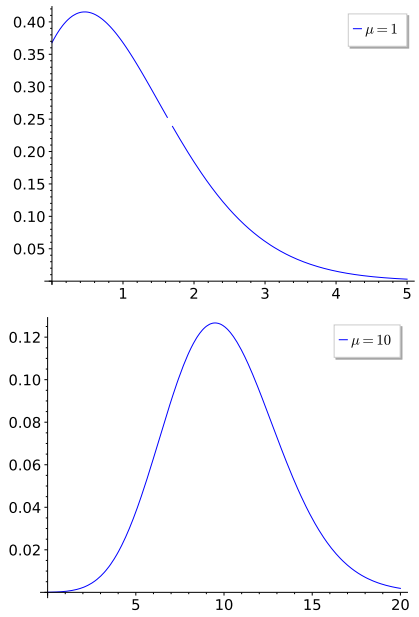
The super nice thing about the Poisson distribution is that it has a super easy expected value and standard deviation. Supposing that X has a Poisson distribution with parameter μ then

$$E(X) = \mu \quad \text{and} \quad \text{SD}(X) = \sqrt{\mu}$$

Example 37.1 It is known that the number of raindrops that fall on a particular square inch of roof in a one-second interval of time is given by the Poisson distribution. Supposing that it has Poisson distribution with parameter 2, what is the variation for the number of raindrops (that fall on a particular square inch of roof in a one-second interval)?

Let's look at a couple of examples of the Poisson distribution with different values of μ.





Notice how, as we plug-in higher μ, we get closer to having a normal distribution. What this means is that we can use our normal approximation for Poisson distributions as well! So if we have a Poisson distribution on a random variable X then

$P(a \leq X \leq b) \approx$

And the skew-normal approximation is given by

$P(X \leq b) \approx$

Another cool thing about the Poisson distribution is that they're relatively easy to sum up. Suppose we have n independent Poisson variables X_1, X_2, ..., X_n with parameters μ_1, μ_2, ..., μ_n respectively. Then we know that $X_1 + X_2 + \cdots + X_n$ is a Poisson random variable with parameter $\mu_1 + \mu_2 + \cdots + \mu_n$.

Example 37.2 Suppose that we're watching a game of bowling where four players are on a team together. The first player has a $\frac{1}{15}$ chance of hitting a strike, the second player has a $\frac{1}{20}$ chance of hitting a strike, the third play has a $\frac{1}{70}$ chance and the final player has a $\frac{1}{10}$ chance. They each will throw the bowling ball 10 times and see how many strikes they get. What is the probability distribution for how many strikes the team will get as a whole? (Use Poisson distribution for each player and assume that each throwing of a ball is independent.)

38 Random Scatters

We're now going to look into a particular way of looking at the Poisson distribution in order to help us tackle a common question. We've stated the problem before, but we'll state it again.

Suppose you're trying to measure how many times a particle hits some surface. You can either think of this as raindrops hitting some sheet, or other particles like dust, molecules, photons, etc. In biology you can think of this as the positions of different cells or organisms on a microscope slide. In astronomy you can think of this as positions of stars in the night sky. In baking, you can think of this as positions of chocolate in a chocolate chip cookie. As you can see, this idea can be thought of in a lot of different ways!

This entire model is dependent on the fact that the number of points in a given area has the Poisson distribution. The same idea holds if we change our dimension from 2 to 3 to 4 or higher! We'll still have the Poisson distribution if looking at the number of points in a given volume.

In order to do this, we'll make two key assumptions:

(1) No point is in the exact same location.

(2) There's an equal chance for the point being anywhere. (Aka, it's random)

The frequency of how many points will occur will be estimated by a constant λ. This gives us the Poisson scatter theorem.

Wikipedia: Poisson Scatter Theorem

Theorem 38.1 (Poisson Scatter Theorem) *Suppose we have a square sheet where there are no points in the exact same location and all the points are randomly placed.*

Let B be a subset of the square and X_B be the number of hits in B.

(1) Then X_B is a Poisson random variable with parameter $\lambda \times area(B)$.

(2) If B_1 and B_2 are disjoint subsets, then X_{B_1} and X_{B_2} are independent events.

This random scatter is called the Poisson scatter with intensity λ. *The intensity λ is the expected number of hits per unit area.*

Conversely, if the above two properties hold, then there are no points in the exact same location and all the points are randomly placed.

Example 38.2 Suppose we live in a fancy house that has a pool, but we're not rich enough to clean the pool during the winter. When summer hits we look at our pool and find a bunch of bacteria growing! In a volume of 1000 drops we find roughly 2000 bacteria (which are all separate from each other and randomly located throughout the water). Loving biology, we take out a single drop and we smear it (uniformly) over the surface of a dish. Since we're totally enamoured with bacteria, we put some food for the bacteria and leave them alone. We come back after a few days and find nice healthy colonies of bacteria on our dish! What is the distribution of the number of colonies over the whole plate and over an area of half the plate? (A colony is basically a large group of bacteria that huddles together after being birthed from the same great great grandbacteria.)

Now suppose that there's a probability of p that each bacterium will die (independently) for whatever reason. What's the distribution of the number of colonies on the whole dish?

This example is actually what is known as the "thinning of a Poisson scatter". If we have a Poisson scatter with intensity λ and each point has a probability of p of staying, then the scatter of points that are kept is a Poisson scatter with intensity λp.

39 Continuous Distributions

We've seen both continuous and discrete distributions so far, but have mainly focused on the discrete. In the discrete case we learned how to handle the expected value and the standard deviation, but now we want to do the same thing for continuous distributions. Remember that the main difference between discrete and continuous distributions is whether we can count them or not. For example, with the binomial distribution, we can go through the cases one at a time. With the normal distribution we couldn't do that! In order to figure out the distribution we had to take the area under the curve. And so now, we get to the part where integration will finally be used.

Let's go back and remember how the probabilities of a normal distribution were determined. We had some function, let's say $f(x)$, and this gave us some bell curve. This bell curve came from looking like the histogram of the binomial distribution. To get the probability at a certain point, the binomial distribution was easy since it was all rectangles. So calculating the area was simple. Calculating the probability at a certain point for the normal distribution was more difficult since we couldn't look at the distribution at a given point since the width was basically zero! So we said we just need to take the integral:

That's it! As long as the function $f(x)$ is continuous, we can take this integral, which is why these distributions are called *continuous distributions*. The function $f(x)$ is called the *probability density*. Other than being continuous, the only other thing we require of a probability density is that the area under the curve must be equal to 1 (or else it's not a distribution). This allows for

Wikipedia: Continuous Distribution
Wikipedia: Probability Density

$$P(\text{everything}) =$$

At this point we might ask how all of the terminology from the

discrete world moves into the continuous world. It turns out, it's almost the same! For example, in the discrete case we had the following formula for expected value:

$$E(X) =$$

If you recall, by making the boxes smaller and smaller, we can turn this into a Riemann sum which leads into integration! So for the continuous case we have:

$$E(X) =$$

To calculate the variance, we have $\text{Var}(X) = E(X^2) - E(X)^2$ and so we also need:

$$E(X^2) =$$

which also gives us the standard deviation $\text{SD}(X) = \sqrt{\text{Var}(X)}$.

Ok, so what about independence? Remember for the discrete case we had the following definition for independence

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

Notice how for the continuous case this formula doesn't help! Since the probability at any point is equal to 0 we have that both sides always equal 0. In other words, if we used this formula, every continuous distribution would be independent, which doesn't make sense. We *won't* be covering how to define independence in the continuous case. If we have time at the end of the year, I'll cover it as a bonus, but for now we will skip it. The main thing to know is that *if* two continuous random variables are independent then all other independence rules hold (like we still have $E(XY) = E(X)E(Y)$ if X and Y are independent and both $E(X)$ and $E(Y)$ are defined and finite).

Week 8

8 July 2021

40 Continuous Uniform Distribution

A random variable X has *continuous uniform distribution on the interval* (a, b) if X has density function $f(x)$ which is constant on the interval (a, b) and 0 everywhere else. The density function is given by:

$$f(x) =$$

Wikipedia: Continuous uniform distribution
Note that the book calls this the "Uniform distribution", but since there is both a discrete and a continuous uniform distribution, we distinguish the two.

Why can't we chose any constant? _____

In this case, the probability function is super easy!

$$P(c \leq X \leq d) =$$

What about the expected value?

That's a little complicated, but there is an easier way! We can scale our continuous uniform distribution to a "standard" length. If you recall from the beginning of the semester, we can rescale this to the interval

$(0,1)$. This is called the *Standard continuous uniform distribution*. It is the continuous uniform distribution on $(0,1)$.

So how do we rescale? Well, first we need to move the entire interval over by a and then we need to scale b down to 1. So we get a standard continuous uniform random variable U where

$$U =$$

Converting this, we get:

$$X =$$

How does this help?

That integral was much easier to solve! What about the variance? Well, we know that

$$E(U^2) =$$

So

41 Normal Distribution

We've used the normal distribution a lot for approximating the binomial distribution. The *standard normal distribution* is the distribution with density function

$$\phi(z) =$$

Remember that taking the integral of this function is notoriously hard, so instead I'll give you integrals to help out:

$$\int_{-\infty}^{\infty} \phi(z)\,\mathrm{dz} = \_\_ \qquad \int_{-\infty}^{\infty} z\phi(z)\,\mathrm{dz} = \_\_ \qquad \int_{-\infty}^{\infty} z^2\phi(z)\,\mathrm{dz} = \_\_$$

What this means is that if Z is a standard normal random variable then $E(Z) = 0$ and $\mathrm{Var}(Z) = 1$.

But what if we want our normal formula where we move things left and right and deviate more than normal? (aka the non-standard versioin) Then we let $X = \mu + \sigma Z$ implying $Z = \frac{X-\mu}{\sigma}$ which should be a formula we've seen before! This gives us a density function

$$\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

With this density function, we have our *normal distribution* The expected value is:

Wikipedia: Normal Distribution

$$E(X) =$$

The variance is:

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

The normal distribution is commonly referred to as the *Gaussian distribution*.

Note that since this integral is super difficult, we will continue to use the Φ approximation that we learned in week 3.

Example 41.1 Suppose that we take repeated measurements of the weight of a standard kilogram. Over time, this will slowly decrease in mass! We know that the weight is a normal distribution with expected value 1 kilogram and standard deviation 20 micrograms. What (approximate) proportion of measurements are correct to within 10 micrograms?

42 Arbitrary continuous distributions

In this section we'll look at two examples of continuous distributions which are not ones that are normally seen. Like this you can see how we can find a distribution from arbitrary density functions and not just the "normal/standard" ones. We do these through examples.

Example 42.1 Suppose that we have a bacterial colony which appears uniformly distributed at random on a circular plate of radius 1. Let R be the random variable which represents the distance from the centre of the plate. What is the probability density, probability (from a to b), expected value and variance of R.

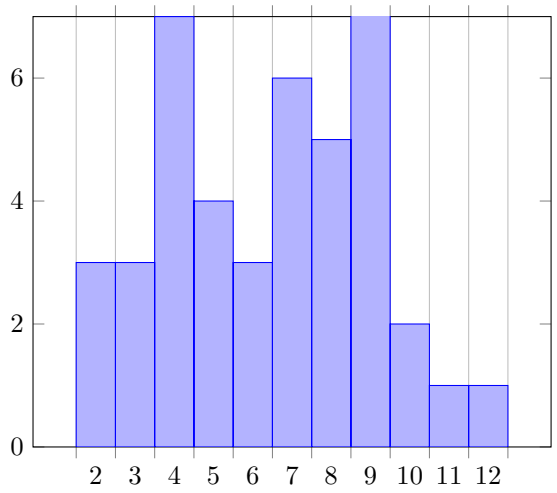
Example 42.2 Let X have the probability density

$$f(x) = \begin{cases} \frac{1}{(1+x)^2} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is $P(X > 3)$ and $E(X)$?

43 Empirical distributions

In this section we'll look at what happens when we take empirical data and try and create a continuous distribution out of it. Say I have done some experiment multiple times and I get some random histogram that looks bizarre.



With the above histogram, we might want to draw a "best-fit" curve. This is normally done by something called "numerical analysis" and is a whole topic in mathematics. We won't get into it to much, but there are ways to approximate this curve. For us, you'll be given this approximation so don't worry about it. (Numerical analysis is actually a real cool class!)

The idea is that if we have a random histogram after some experiments, we can ask a numerical analyst to look at the data and create a best-fit function. We can then take that function, say $f(x)$, and create a probability out of it if we add the stipulation that the area under the curve must be 1. This distribution is given by

$$P(a,b) \approx$$

We can actually use indicator functions to help us compute averages from our empirical data. If we let $I_{(a,b)}(x)$ be the indicator function

$$I_{(a,b)}(x) =$$

then we get a way to look at our probabilities. Looking at our histogram we have

$$P(a,b) =$$

where x_i is inside some list of numbers $(x_1, x_2, \ldots, x_n)$ from our empirical data. This is basically keeping track of how many of our boxes actually occur. If we combine our two ideas, we end up with

$$P(a,b) \approx$$

This is true since x is 0 outside of (a,b) and so we can make this assertion.

This might seem like a complicated thing to do, because we're basically multiplying by 1, but it actually gives us a powerful tool. What we basically have shown is that

We can generalize this to give us the *integral approximation for averages*. To generalize we just use an arbitrary function $g(x)$ instead of the indicator function.

Notice how both of these functions are giving us the expected value. The left hand side is the expected value of $g(X)$ where X is picked at random from the list $(x_1, x_2, \ldots, x_n)$ (aka discrete version). The right hand side is the expected value of $g(X)$ for a random variable X with density $f(x)$.

One thing, right off the bat that we can use this for is moments. If we let $g(x)$ denote the kth moment, *i.e.*, _____, then we have

The left hand side is known as the *kth moment of the empirical distribution*. The right hand side is known as the *kth moment of the theoretical distribution*.

We never defined the empirical distribution, but you can read more about it on wiki.
Wikipedia: Empirical Distribution

When $k = 1$ and $k = 2$ we have enough information to calculate the expected value and the variance.

But how good is our approximation? To be honest, it's definitely a little complicated. The following is Chebychev's inequality for any $\varepsilon > 0$:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{dx}\right| > \varepsilon\right) \leq \frac{\mathrm{Var}\left(g(X)\right)}{n\varepsilon^2}$$

I won't solve this or show that it's true. There's more information in the book if you'd like, but the main thing to know here is that we are able to calculate how close the approximation is. Another thing to notice is that if n is very large compared to the variance, then the approximation gets better and better.

Why do we care about all of this? Because integrals are hard! What this is basically telling us is that if we have a difficult integral, we can estimate the integral by a sum! If we take a uniform set of n values, we can approximate our integral using the left hand side. This method is

Wikipedia: Monte Carlo method

known as the *Monte Carlo method* and is used heavily in physics.

44 Exponential Distribution

Next up, we'll look at distributions where something changes over time. You can think of this as the time it takes for a computer process to be completed, the time it takes for atoms to decay, the time it takes for organisms to evolve, the time it takes for water to boil, etc. Usually time starts from now and goes on into the future. Mathematically we can think of "now" as 0 and the far, far future as ∞. So, for example, if I want to think about the probability of something happening between 1 and 2 days from now, then if I let T be the random variable which calculates the number of days, then I'm asking for $P(1 \leq T \leq 2)$. Since we're talking about continuous distributions this week, you can expect this to be equal to an integral:

$$P(1 \leq T \leq 2) =$$

If we want to think of something just goes on forever starting from a certain time d, then we can calculate

$$P(d \leq T) =$$

What this means is we can break down $P(a \leq T \leq b)$ into the following

$$P(a \leq T \leq b) =$$

What function we chose to put for $f(t)$ changes based on the model or the question we are asking. There are many different distributions that can be used with respect to time in this way, but the main one we will focus on in this section is the *exponential distribution*.

The *exponential distribution with rate* λ is the continuous distribution where

$$f(t) =$$

Wikipedia: Exponential distribution

for $t \geq 0$.

What does this distribution look like?

$$P(a \leq T \leq b) =$$

Notice that if $a = 0$ and $b = \infty$, then we get that the entire area is equal to 1.

But look at what this formula implies! If I kept $b = \infty$ and I let a be anything then we have

$$P(a \leq T) =$$

That is so cool! It's actually a super easy formula for us to remember. The nice thing about this is that the expected value and the standard deviation are also really nice numbers:

$$E(T) = \qquad \text{and } \mathrm{SD}(T) =$$

So the formula is cool and all, but what does this λ mean? λ is basically the value of the instantaneous "bye bye" rate (aka death rate). In other words, it keeps track of how much of a certain thing goes away after a certain amount of time. (Depends on what your "unit" of time is)

Example 44.1 Suppose that you just bought a new charging cable for your phone. You know that, on average, the cable will last you 30 days and you are told that the distribution is exponential. What is the probability that the cable will be working after 15 days?

45 Poisson point process

Remember how the normal/Poisson distribution(s) could be seen as the continuous version(s) of the binomial distribution? It turns out that the exponential distribution is just the continuous version of the geometric distribution! For the geometric distribution we looked over the numbers $\{0, 1, 2, 3, \ldots\}$, but in the exponential we look at all non-negative numbers. For a geometric random variable X we had that $E(X) = \frac{1}{p}$ where p is the probability. If our probability is small enough, then $E(X)$ is very large! We can then take this and rescale X by the expected value:

$$\frac{X}{E(X)} =$$

Then, recalling that $P(X > n) = (1-p)^n$ in the geometric case, we have

$P(pX > t) =$

This last approximation is coming from the fact that $(1-p) \approx e^{-p}$ for really small p. For example, if $p = 0$:

There's also a way to look at the exponential distribution through the lens of the Poisson distribution. When we had first looked at the binomial distribution, we were keeping track of how many successes we have in n trials. If we had k successes then we had

for the binomial distribution. But, there's another way to think about this. Instead we can look at *when* certain successes occur and look at the gaps in between them.

As a quick example, let's see what I mean about the two ways of looking at this. Suppose we have 10 trials and 2 successes. As an example, we can think of the black dots as successes and the white dots as failures:

(1) In the first way, we look at it by the number of successes that happen altogether and where these successes might occur. In our example, this is represented by the black circles being in the 1st and the 7th positions. This is where the $\binom{10}{2}$ number is coming from.

(2) In the second way, we look at the gaps *between* successes. In other words, with our example, we see there are no failures, followed by 1 success, followed by 5 failures, followed by 1 success, followed by 3 failures. So we can represent this as $(0, 5, 3)$. In other words, we can think of this as the ways to add up any three (non-negative) numbers a, b, c such that their sum is $n - k = 10 - 2 = 8$:

$$0 + 5 + 3 = 8$$

Notice that this first way of looking at things is the binomial distribution. On the other hand, looking at things in the second way is the geometric distribution (since we ask for "how long do I have to wait before something happens"). These two ways of looking at things give a description of what's called the *Poisson point process with rate λ*.

Wikipedia: Poisson point process
The book calls this the "Poisson arrival process".

(1) **Points/Arrivals:** If we let I be a fixed (time) interval of length t then we can let $N(I)$ be the number of "successes" (usually called *arrivals*). This gives us a Poisson distribution with parameter λt. In this case, λ is the number of successes we expect to see per unit of time.

(2) **Counts/Times:** Alternatively, we can start from the beginning ($t = 0$) and count the number of success we get as our time increases. This gives us an exponential distribution with expected value $\frac{1}{\lambda}$.

These two ways of looking at things are equivalent.

Example 45.1 The standard example in this case is to look at phone calls, but we'll use dms instead. Suppose that you get roughly 3 dms per minute (look who's popular!).

Using the first way of looking at the Poisson point process, we can look at this as a Poisson distribution. Since we're talking about "per minute" our unit of time is __ minute(s). If we let our interval I be from 2 minutes in the future to 6 minutes in the future, then we know that $t =$ __. So then our Poisson distribution parameter is given by _____.

Using the second way of looking at the Poisson point process, we can find an exponential distribution. Since we get 3 dms per minute, that means our average wait time between dms is $\frac{1}{3}$ of a minute. This means, the expected value is $\frac{1}{3} =$ ___. In other words, we have an exponential distribution with rate _____.

What is the probability that you get 0 dms in the first four minutes ($t = 0$ to $t = 4$)?

What is the probability that we get our first dm after 4 minutes ($t = 4$)?

What is the probability that you get no dms in the first minute and then at most 3 dms in the second minute?

What is the probability that you get no dms in the first minute and the wait time between your second and third dm is more than 2 minutes?

What is the probability that your fifth dm takes more than 4 minutes to arrive?

46 Erlang Distribution

Our next distribution uses the Poisson point process to create a new distribution. If we look at the second definition of the Poisson point process (where we look at the time between successes), then we can keep track of the time of the kth arrival after time 0. Let T_i be the time it takes for a success between the $i-1$st success and the ith success, *i.e.*, we assume that T_i has an exponential distribution. So T_1 is how long it takes from 0 until the first success, T_2 is how long it takes from T_1 until second success, etc. Then we let $G_r = T_1 + T_2 + \ldots + T_r$, *i.e.*, how long it takes from 0 until the rth success. The distribution $P(G_r)$ is known as the *Erlang distribution with parameters λ and r*. In order to distinguish the two parameters, λ is called the *rate parameter* and r is

Wikipedia: Erlang distribution
Note that the book calls this a "Gamma distribution".

known as the *shape parameter*.

The probability density function for an Erlang distribution is given by the function

$$f(x) =$$

Note that it is dependent on the rate λ and the parameter r. Also, note that r *must* be an integer.

Let X be an Erlang random variable with rate λ and parameter r. Then

$$E(X) = \_\_\_\_ \quad \text{and} \quad \text{SD}(X) = \_\_\_\_\_$$

and the tail sum is given by

$$P(X > x) = \sum_{k=0}^{r-1} e^{-\lambda x} \frac{\lambda^k x^k}{k!}$$

Example 46.1 Suppose that it's winter in Toronto and we know that, on average, it snows once every 10 days. Furthermore, suppose that "it snowing" is an exponential distribution. (In other words, the days it snows are exponentially distributed). What is the expected time for the next four snow days to occur?

What is the probability that the next four snow days will occur next week? (7 days from now until 14 days from now)

46.1 Gamma Distribution

In our final distribution for this week we look at the Gamma distribution. If we look at the Erlang distribution we notice that whenever we have G_r we required that r be an integer since it's counting the number of time something happens. If let r be any real number with the same density function as the Erlang distribution, then we get the Gamma distribution. In other words, the *Gamma distribution with parameters λ and r* is the distribution with density function

$$f(x) =$$

Wikipedia: Gamma distribution

where r is any positive number. In order to distinguish the two parameters, λ is called the *rate parameter* and r is known as the *shape parameter*. Note that if k is an integer, then the Gamma distribution is just the Erlang distribution.

Week 9

15 July 2021

Week 10

22 July 2021

47 Change of variable

We've already seen a few times where we changed variables. The biggest and main example was when we altered a normal distribution X to a standard normal distribution Y. This was done using the formula $Y =$ _____. We continue this idea by using formulas $Y = f(X)$ in order to change one variable into another, but in a continuous sense.

For continuous distributions remember that $P(a \leq X \leq b) =$ _____. Say we then have some new distribution Y such that $Y = g(X)$ where g is some function. The random variable Y is going to have some density function $f_Y(y)$ which can be written in terms of $f_X(x)$ and $\frac{\mathrm{dy}}{\mathrm{dx}}$. How do we convert in between these two?

47.1 Linear functions

The first thing we'll do is look at simple functions, much like $Y = \frac{X-\mu}{\sigma}$. By "simple functions", I really just mean linear functions: _____. In this case we have _____ implying _____. This just scales the length of every interval, which is exactly what happened when we converted a normal distribution into a standard normal distribution.

Let's look at what happens for a particular distribution.

Example 47.1 Suppose that X is a uniform distribution random variable with density function

$$f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Let $Y = aX + b$ for some $a > 0$. What is $f_Y(y)$?

What if $a < 0$?

What this boils down to is the following formula:

$$f_{aX+b}(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

This is true for *any* distribution.

Notice how this is basically what we did for normal distributions! We had the function $Y = \frac{X-\mu}{\sigma}$ implying that $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$. In this case Y was the standard normal distribution (with mean 0 and standard deviation 1) and X was the normal distribution (with mean μ and standard deviation σ). If we let $\phi(x)$ denote the function $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ then, converting to y we have

$$\phi(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

47.2 One-to-one change of variable for densities

Now we suppose that X is a random variable with density function $f_X(x)$ on the interval (a, b). We let Y be another random variable such that $Y = g(X)$ where g is strictly increasing (or decreasing) on the interval (a, b). From our discussion in the previous section, Y is defined on either

the interval $(g(a), g(b))$ or $(g(b), g(a))$. We now want to calculate $f_Y(y)$ for y in one of these intervals. Thinking about our probability functions, we have that

$$f_Y(y)\,\mathrm{dy} = f_X(x)\,\mathrm{dx} \Rightarrow f_Y(y) = \frac{f_X(x)}{|\mathrm{dy}\,/\,\mathrm{dx}|}$$

where $y = g(x)$.

Example 47.2 Suppose that X has exponential density $f_X(x) = e^{-x}$, $x > 0$. What is the density of $Y = \sqrt{X}$?

Example 47.3 This time, suppose that X has a uniform distribution on $(0, 1)$. Let $Y = \frac{-\log(X)}{\lambda}$ where $\lambda > 0$. What is the density of Y?

What this means is that we can use functions to change a random variable of one distribution into a random variable of another distribution! This change of distribution is *dependent* on the function g. In other words if X and Y have the same distribution, then $g(X)$ and $g(Y)$ have the same distribution.

47.3 Many-to-one change of variable for densities

What about if we have a function that isn't just increasing/deceasing? We can split it up into multiple little parts that handle the increasing/decreasing part, one at a time. In other words, we have

$$f_Y(y) = \sum_{\{x \mid g(x)=y\}} \frac{f_X(x)}{|\mathrm{dy} \,/\, \mathrm{dx}|}$$

Best way to do this is through an example.

Example 47.4 Suppose that X is a random variable an let $Y = X^2$. What is the probability density function for Y?

47.4 Other change of variables

In this section, we're just going to look at two other examples of change of variables that aren't dependent on "restrictions" in order to see how to do this using any arbitrary function g.

Example 47.5 Suppose you want to cut a cake. You grab your unit cake (*i.e.*, radius is equal to 1) and you randomly choose a point on the side to start on. (By random here I mean "uniformly at random") Let X be the x-coordinate of where you begin to cut the cake, what is the probability density and the expected value of X?

Before you make some weird cut, your cousin looks at you and asks for your knife before you cut yourself. They tell you, you can't just cut your cake from any side, it has to be from the top half. In other words, we let $Y = |X|$. What is the probability density function of Y and the expected value of Y?

Example 47.6 Let's do another example, but on a sphere this time.

Let Θ be the latitude, between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ of a point chosen uniformly at random on the surface of a unit sphere. First, what is the probability density of Θ?

Let Y be the vertical coordinate of the point on the sphere between -1 and 1. What is the probability density function of Y?

48 Cumulative Distribution Functions

With a lot of the continuous distribution functions, we normally looked at probabilities in a range $P(a \leq X \leq b)$, but there's another way we can look at these distributions. Let X be a random variable, then the *cumulative distribution function* is the function

Wikipedia: Cumulative distribution function

We've actually already seen an example of this. Remember that for the standard normal distribution, we have that $P(X \leq x) =$ _____. In other words, _____. Using this, we know that we can take this a little further! Recall that we also stated that $P(a \leq X \leq b) =$ _____. We can actually do this with any distribution.

Notice how for the standard normal distribution I had $P(a \leq X \leq b)$ and above I wrote $P(a < X \leq b)$. Why is the first one a $\leq$ and the second one a $<$? The reason we're allowed to use $\leq$ for the standard normal distribution is because it's continuous

To make this more precise, we say that a distribution is *continuous* if its cumulative distribution function is a continuous function. It can be shown that if a distribution is continuous, then _____ for every x. Since that probability is 0, changing $<$ to $\leq$ doesn't change the probability. In other words we have

$$P(X > x) = 1 - F_X(x)$$

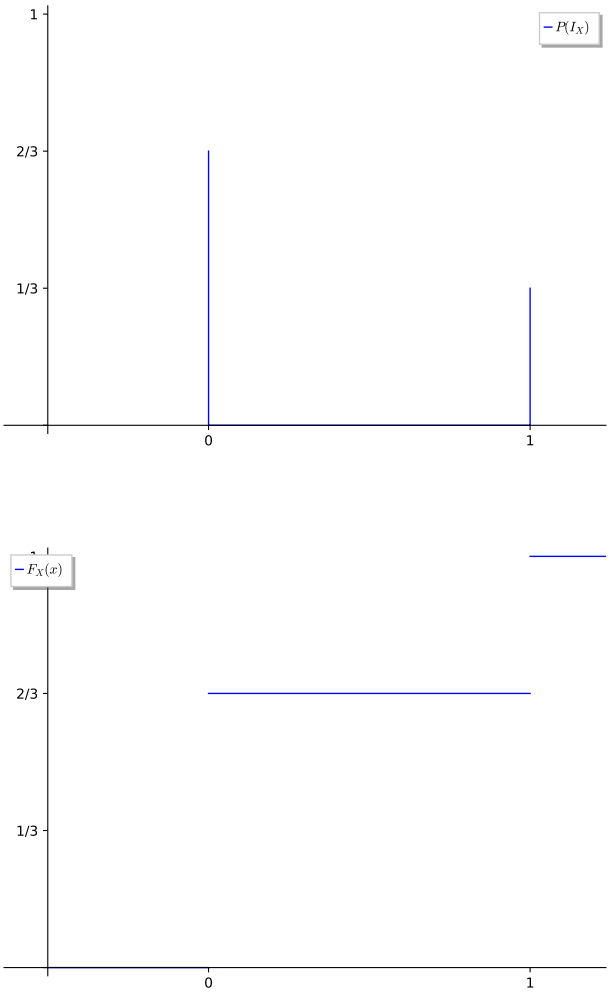
for *every* distribution and

$$P(X \geq x) = 1 - F_X(x)$$

is *only* true for continuous distributions. In fact, a probability distribution is determined by its cumulative distribution function.

48.1 Discrete Case

Suppose we have an unfair coin which is heads 1/3 of the time and tails 2/3 of a time. Let X be the flip of the coin. In order to describe this using numbers, we'll use an indicator function. So let I_X be the indicator function for X, *i.e.*, $I_X = 1$ if we flip a heads and $I_X = 0$ if we flip a tails. By the above we have $P(I_X = 1) = P(X = \text{ heads}) = \frac{1}{3}$. Then we have



It's easy to see that this is not a continuous function and so this distribution is not continuous. Usually a discrete distribution looks like a staircase, such as above. This should also show why we can't have $\leq$.

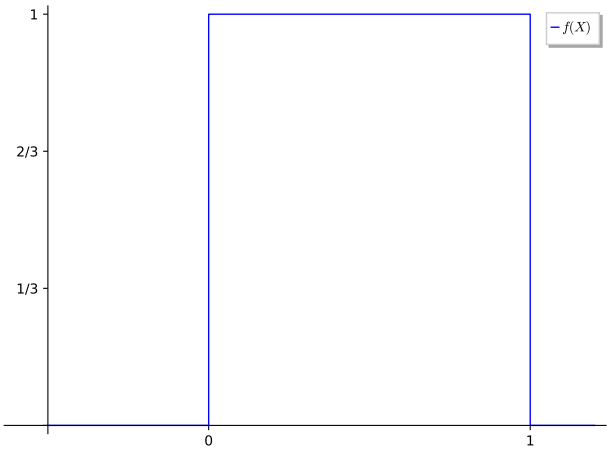
48.2 Continuous case

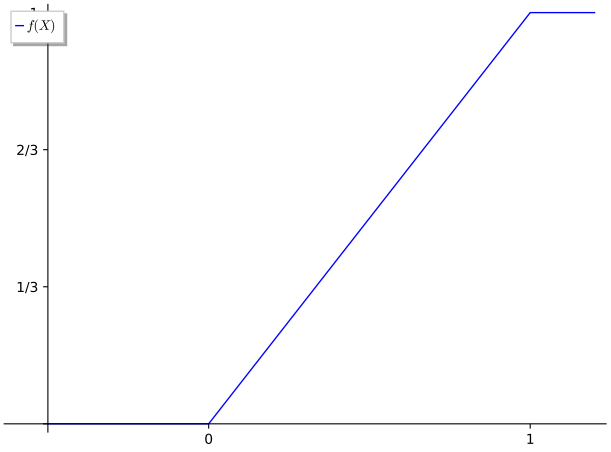
If our distribution is continuous, then we already know what $F_X(x)$ is:

$$F_X(x) =$$

In other words, it's just the _____ of the probability density function.

Example 48.1 Let X be a random variable which has a standard uniform distribution. What is its cumulative distribution function?


1
2/3
1/3
0
1
f(X)


f(X)
2/3
1/3
0
1

49 Maximum and Minimum of Independent Random Variables

One function of random variables that we've been ignoring so far are the max and min functions. Say for example I have n random variables $X_1, X_2, \ldots, X_n$ which are all independent. Then

$$X_{\max} = \max(X_1, X_2, \ldots, X_n) \quad X_{\min} = \min(X_1, X_2, \ldots, X_n)$$

Let's first look at $X_{\max}$. Notice that if $X_{\max} \leq x$ then that implies *every* $X_i \leq x$. Let $F_{\max}(x)$ be the CDF of $X_{\max}$. Then

$$\begin{aligned} F_{\max}(x) &= P(X_{\max} \leq x) \\ &= P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) \\ &= P(X_1 \leq x) \cdot P(X_2 \leq x) \cdot \ldots \cdot P(X_n \leq x) \\ &= F_{X_1}(x) \cdot F_{X_2}(x) \cdot \ldots \cdot F_{X_n}(x) \end{aligned}$$

Letting $F_{\min}(x)$ be the CDF of $X_{\min}$. Then we similarly have

$$\begin{aligned} F_{\min}(x) &= P(X_{\min} \leq x) \\ &= 1 - P(X_{\min} > x) \\ &= 1 - P(X_1 > x, X_2 > x, \ldots, X_n > x) \\ &= 1 - P(X_1 > x) \cdot P(X_2 > x) \cdot \ldots \cdot P(X_n > x) \\ &= 1 - ((1 - F_{X_1}(x)) \cdot (1 - F_{X_2}(x)) \cdot \ldots \cdot (1 - F_{X_n}(x))) \end{aligned}$$

Example 49.1 Let $X_1, X_2, \ldots, X_n$ be independent random variables where each X_i has exponential distribution with rate λ_i. What is the distribution of $X_{\min} = \min(X_1, X_2, \ldots, X_n)$?

50 Quantile function

Let X be a random variable of some distribution and let $F_X(x) =$ _____ be the cumulative distribution function. So far we've been calculating $F_X(x)$ given some value x, but sometimes we want to do the inverse. Say for example we're given _____ for some variable p, what value of x satisfies this equality? To do this, we need to take the inverse: _____.

Wikipedia: Quantile function
Note that the book calls this the "inverse distribution function".

Notice that $F_X^{-1}(p)$ is just the inverse of $F_X(x)$ and is therefore called the *inverse cumulative distribution function*, or more commonly the *quantile function*. This function is only defined when $0 < p < 1$ and the solution is normally called the *pth quantile*. For example, _____ is known as the $\frac{1}{12}$-quantile.

Some common names for these quantiles are:

(1) *Quartile*: $p =$ ___

(2) *Median*: $p =$ ___

(3) *Decile*: $p =$ ____

(4) *Percentile*: $p =$ _____

Example 50.1 Let X be a random variable of an exponential distribution with parameter λ. Find a formula for the p-th quantile.

Week 11

29 July 2021

51 (Discrete) Conditional Distributions

Earlier in the term, we had gone over conditional distributions in the discrete case, but at that time we'd only seen a couple of discrete distributions. Now that we know more distributions, we'll review some of the topics on conditional distributions and apply it to other distributions. Remember that if we have two random variables X and Y, then the conditional probability of Y being y when X is x is given by

Wikipedia: Conditional probability distribution

This is known as the *conditional distribution of Y given $X = x$*.

Recalling that with random variables we use "," instead of "$\cap$" gives us the normal multiplication rule that we've used before:

We can also break this down into the law of total probability:

Example 51.1 Suppose I roll a fair six-sided die. Then, I look at the number rolled, and I write down all the divisors of that number. For example, if I roll a 6, then the divisors of 6 are $1, 2, 3, 6$. I then roll my six-sided die repeatedly until I roll one of the divisors. What is the distribution on the number of times I need to roll the six-sided die the second time around?

52 (Discrete) Conditional Expectation

A question we have yet to answer is: if one thing is conditioned on another thing (like in the previous section), how do we calculate the expected value? This is actually much easier than one might think. The *conditional expectation of a random variable Y on an event A* is

Wikipedia: Conditional expectation

Example 52.1 Let Y be the number of heads in four tosses of a fair coin. Suppose we know we received at most two heads. What is the expected value of Y?

There are other rules which translate nicely for conditional expectation.

For example, *linearity* still holds:

Wikipedia: Law of total expectation

The *law of total expectation* holds:

This last formula makes us wonder if we can do a random variable *given* another random variable. From above, we know that

$$E(Y \mid X = x) =$$

But what if we want $E(Y \mid X)$?

In this case $E(Y \mid X)$ is defined as *conditional expectation of Y given X* and is a function of x. In other words $E(Y \mid X)$ is a function f such that:

In this case, $E(Y \mid X)$ can be considered either a formula/function (like above) or as a random variable itself. Why is it a random variable? Since the function is over all of x and so we get a distribution!

But this makes things a little weird. For example, it makes sense to ask what $E(E(Y \mid X))$ is equal to since $E(Y \mid X)$ is a random variable. What does it equal?

This gives us another way to look at the law of total expectation:

Example 52.2 Let's look at a previous example. Suppose I roll a fair six-sided die. Then, I look at the number rolled, and I write down all the divisors of that number. I then roll my six-sided die repeatedly until I roll one of the divisors. Let X be the number rolled by the die and let Y be the number of rolls needed to get a divisor of X.

What is the conditional expectation of Y given $X = x$?

What is the expected value of Y?

Example 52.3 Let S_n denote the number of successes of n independent trials with probability p of success for each trial. What is $E(S_m \mid S_n = k)$ for $m \leq n$?

53 Functions in (discrete) conditional expectation

Intuitively if we're *given* $X = x$ we want to replace all X with x since that's what we know. So if we were to have a function $g(X, Y)$ and we're given $X = x$ we want to do something like $g(x, Y)$. It turns out this is perfectly ok!

This is super useful for when we do functions with multiple variables.

This implies that

Likewise, if we look at a linear equation

implying that

So this helps us expand conditional expectations much more simply!

Example 53.1 If X and Y are independent, find $E(X+Y \mid X=x)$.

54 Covariance

Remember from halfway through the course, we had mentioned that if two random variables X and Y are independent then

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

We can actually take this and break it down for any two random variables, not just if they're independent.

Wikipedia: Covariance

This final component _____ is known as the *covariance* and is denoted by _____.

Some quick things to notice:

$\text{Cov}(X, X) =$

If X and Y are independent then $\text{Cov}(X, Y) =$ __. Note that the converse is *not* true. In other words, if $\text{Cov}(X, Y) =$ __ then we don't know if X and Y are independent or not!

54.1 Indicators

Before we look at an example, we'll look at what happens when our random variables are indicator functions. Let X_A be an indicator function for an event A and let X_B be an indicator function for an event B. We should also note that $X_A \cdot X_B$ is the indicator function for event A and B. Then we have

$E(X_A) =$ _____, $E(X_B) =$ _____, $E(X_A X_B) =$ _____

Implying

$\text{Cov}(X_A, X_B) =$

In the case of indicator functions (and it is one of the few cases this is true) we have the following:

$\text{Cov}(X_A, X_B) > 0 \iff P(A \cap B) > P(A)(P(B);$ A and B are *positively dependent*
$\text{Cov}(X_A, X_B) = 0 \iff P(A \cap B) = P(A)(P(B);$ A and B are *independent*
$\text{Cov}(X_A, X_B) < 0 \iff P(A \cap B) < P(A)(P(B);$ A and B are *negatively dependent*

But what does this mean? If A and B are positively dependent, it just means that the chance of A given B is greater than the chance of A happening. This is true by using our conditional probability. Recall that

$$P(A \mid B) =$$

So this implies

$$P(A \cap B) > P(A)P(B) \Rightarrow$$

Similarly we can swap A and B and also look at things from the negative perspective.

Example 54.1 Consider a box with 6 balls in it, 3 black and 3 white. Let B_1 and B_2 denote pulling out a black ball in the first and second pull. Let W_1 and W_2 denote pulling out a white ball in the first and second pull.

What are the dependencies of B_1, B_2 and B_1, W_2 if we do a sampling with and without replacement?

55 Correlation

What the indicator example told us was that the sign of the covariance tells us some information about the dependence between two random variables. Generally, if $\text{Cov}(X, Y)$ is positive, then above average values of X tend to be associated with above average values of Y and below average values of X are associated with below average values of Y. Likewise, if $\text{Cov}(X, Y)$ is negative, then above average values of X tend to be associated with below average values of Y and below average values of X are associated with above average values of Y. If $\text{Cov}(X, Y) = 0$ then there's no real association and we can't really say much. Although the sign gives us information about how two random variables are associated with one another, the magnitude doesn't tell us much unless we dilute it. Diluting the magnitude will give us the correlation between two random variables which is what we define next.

Wikipedia: Correlation

The *correlation* between two random variables X and Y is given by

As long as X and Y are not constant, the denominator is always positive and so Corr and Cov have the same sign, but the magnitude changes.

We say that two variables are *uncorrelated* if one of the following three (equivalent statements) hold

(1) _____

(2) _____

(3) _____

This implies if two random variables are independent they are uncorrelated (but if two random variables are uncorrelated, they're not necessarily independent).

It can be shown that correlations are always between -1 and 1:

$$-1 \leq \text{Corr}(X, Y) \leq 1$$

We won't prove this, but you can look it up in the book if you'd like.

Example 55.1 Suppose we're playing roulette where there are 18 red numbers, 18 black numbers and 2 green numbers. Let N_R be the number of reds that appear in n spins and let N_B be the number of blacks that appear in n spins. What is the correlation between N_R and N_B?

56 Variance of a sum of many variables

What if we don't just have a sum of two variables, but instead we have the sum of multiple variables? We won't show the proof, but it turns

out

$$\text{Var}\left(\sum_{k=1}^{n} X_k\right) =$$

Example 56.1 Suppose that we're looking at the height of the population of Canada. Let X_i be the height of the ith person. Let μ denote the average height and let σ denote the standard deviation of height. Let $\bar{X_n} = \frac{X_1+\ldots+X_n}{n}$.

Calculate the expected value and standard deviation of $\bar{X_n}$ if the n individuals are chosen with replacement and also calculate them if the individuals are chosen without replacement.

57 Bilinearity of Covariance

Remember that when we had expectation and we did the following

$$E(X+Y) = \underline{\qquad\qquad\qquad\qquad}$$

we called this the *linearity of expectation*. This comes from the word "linear" which has a very unique meaning in math. A map f is called *linear* if $f(x+y) = f(x)+f(y)$ and $f(ax) = af(x)$ for some constant a. This is true for expectation, but it's true for a *lot* of things. For example if $f(x) =$ ____ then

Wikipedia: Linear map

We want to expand this to maps which have two terms (like Cov). Since there are two terms, if a map is linear in both terms, then it's called *bilinear*. For example, $f(x,y) =$ _____ is bilinear. To do this, we check that each of the terms are linear:

Wikipedia: Bilinear map

Note that, we need this to be true for *both* terms! For example if we let $f(x,y) =$ ___ then f is linear in the first term, but not in the second:

It turns out that covariance is bilinear!

Appendix A

Book vs. The world

Book lingo	Standard lingo
Outcome space	Sample space
AB	$A \cap B$
Uniform distribution on a finite set	Discrete uniform distribution
Uniform (a, b) distribution	Continuous uniform distribution
Rule of average conditional probabilities	Law of total probability
Binomial (n, p) distribution	Binomial distribution
Mean	Expected value
Dummy variable	Variable
Same distribution	Equal in distribution
Addition rule for expected value	Linearity of expectation
Poisson arrival process	Poisson point process
Gamma distribution	Erlang distribution
Inverse distribution function	Quantile function
Rule of average conditional expectations	Law of total expectation

References

[1] Jim Pitman. *Probability*. New York: Springer, 1993. ISBN: 9781461243748.

Index